Universidad Zaragoza
1542

Trabajo Fin de Grado

Pipeline para la extracción de texto estructurado de Guías de Práctica Clínica en formato PDF

Pipeline for the extraction of structured text from Clinical Practice Guidelines in PDF format

Autor

Víctor Miguel Peñasco Estívalez

Director

Carlos Telleria Orriols

ESCUELA DE INGENIERÍA Y ARQUITECTURA

2020

RESUMEN

En este Trabajo Fin de Grado se ha desarrollado una herramienta en Python que permite extraer el texto de un documento en formato PDF, procesarlo de manera que se pueda diferenciar entre contenido estándar y títulos de apartado y subapartado, y presentarlo en formato JSON, de forma que cada nodo representa un título o párrafo de texto, preservando la estructura original del documento.

La herramienta consiste en un pipeline de procesamiento en el que primero se realiza una extracción a HTML del documento utilizando la librería open source PDFMiner, se continúa con una depuración del texto en este formato, posteriormente se realiza una conversión a MarkDown para facilitar la creación de párrafos y corrección de defectos mediante expresiones regulares, y finalmente se reconstruye la estructura original mediante una jerarquía de nodos y subnodos de texto en JSON.

Debido a las ilimitadas formas en las que se puede construir un documento PDF, este trabajo se ha enfocado en ofrecer el mejor resultado posible para Guías de Práctica Clínica en dicho formato, y siempre con el objetivo de maximizar la capacidad de realizar posteriormente análisis avanzados de procesamiento de lenguaje natural.

Para comprobar la utilidad y posible continuidad de este trabajo, se ha creado una prueba de concepto de uno de los casos de uso de la herramienta. Se ha desarrollado un sistema que, de forma automática, alimenta un chatbot construido mediante la librería open source Snips NLU, que ofrece una interacción en la que da respuesta a preguntas sobre diversos problemas de salud que son contestadas en Guías de Práctica Clínica. Al transformar las guías a formato estructurado JSON, la detección automatizada de preguntas en títulos de apartados y la selección de su contenido como respuesta se convierte una tarea prácticamente trivial.

Índice

1. Introducción y objetivos **1**

2. Comparación de librerías **5**

2.1. PDFMiner 5

2.1.1. Extracción en formato Texto 5

2.1.2. Extracción en formato Tag Extraction 6

2.1.3. Extracción en formato HTML 6

2.1.4. Extracción en formato XML 7

2.2. PDFPlumber 7

2.3. Slate 7

2.4. PDFlib TET 8

2.5. PyPDF2 9

2.6. Línea de comandos 9

2.7. Conclusiones de la comparación 10

3. Componentes del pipeline **11**

3.1. Esquema general 11

3.2. Extracción a HTML 11

3.3. Procesado de HTML 13

3.3.1. Eliminación de elementos gráficos 13

3.3.2. Eliminación de encabezados y pies de página 13

3.3.3. Eliminación de texto vertical o rotado 15

3.3.4. Ordenación de contenedores de texto 16

3.4. Conversión a MarkDown 17

3.4.1. Análisis de tamaños de fuente 17

3.4.2. Conversión 20

3.5. Procesado de MarkDown 21

3.6. Conversión a JSON 24

4. Prueba de concepto de un caso de uso: chatbot **27**
4.1. Versión básica con Dialogflow 27
4.2. Versión extendida con Snips NLU 28

5. Metodología **31**

6. Conclusiones **35**

7. Bibliografía **37**

Lista de Figuras **39**

Lista de Tablas **41**

Anexos **42**

A. Ejemplo de interacción con el chatbot **45**

B. Ejemplo de extracción y procesado **53**

Capítulo 1

Introducción y objetivos

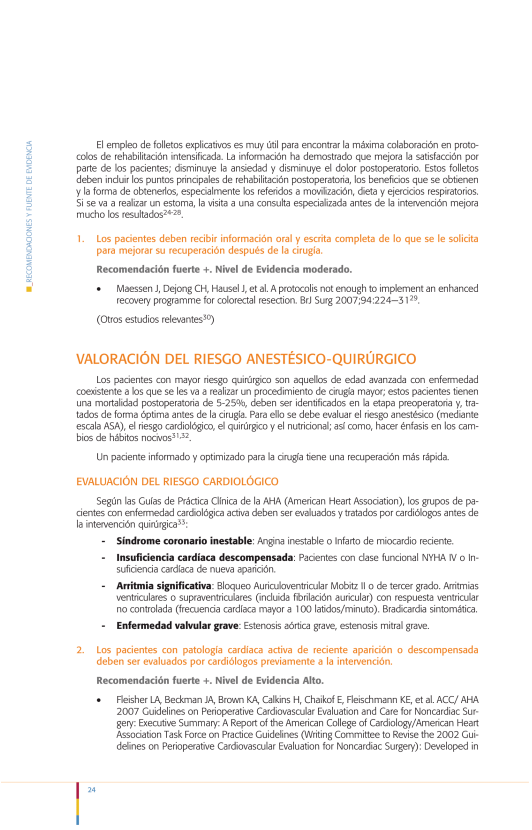
PDF se ha convertido en un estándar global para el almacenamiento e intercambio de información. Una de sus propiedades más interesantes es su capacidad para conservar el aspecto y estructura del documento, independientemente de la plataforma utilizada. Esto lo convierte en un formato excelente cuando el objetivo es la lectura de documentos por parte de personas. En cambio, cuando se quiere que una máquina procese el contenido de estos documentos, surgen varios problemas. La mayoría de programas de extracción de texto de PDFs permite obtener de forma sencilla los bloques de texto encontrados (en su mayoría líneas de texto). Este resultado tiene el inconveniente de perder toda información de estructura que un ser humano puede percibir con la simple lectura del documento, tal y como se puede observar en la Figura 1.1.

En este trabajo se ha desarrollado una herramienta que convierte un documento PDF a un formato estructurado como es JSON, intentando mantener, de la manera más fiel posible, la estructura original de su contenido.

Además, es de especial interés que el software desarrollado sea distribuido de forma libre, para que su aplicación no quede reducida únicamente a este Trabajo Fin de Grado. Es por esto que una de las prioridades en su elaboración es la utilización de librerías y materiales de terceros open source. Por esta misma razón, el código desarrollado en este proyecto está disponible en Github (`https://github.com/vpec/lyrapdf`).

Las Guías de Práctica Clínica [1] son documentos escritos por expertos en distintas patologías y trastornos de salud, en los que se recoge toda la evidencia científica sobre una patología concreta, junto con recomendaciones concretas para su diagnóstico y tratamiento, dirigidas tanto a profesionales como a pacientes. Todas tienen una estructura similar, ya que existe una metodología específica para la elaboración de estas guías, y están disponibles públicamente en formato PDF.

Aunque cualquier documento PDF puede ser procesado con el software desarrollado en este proyecto, los documentos que se han usado como ficheros de prueba son las Guías de Práctica Clínica, de forma que se ha procurado que el pipeline de procesamiento



RECOMENDACIONES Y FUENTE DE EVIDENCIA

El empleo de folletos explicativos es muy útil para encontrar la máxima colaboración en protocolos de rehabilitación intensificada. La información ha demostrado que mejora la satisfacción por parte de los pacientes; disminuye la ansiedad y disminuye el dolor postoperatorio. Estos folletos deben incluir los puntos principales de rehabilitación postoperatoria, los beneficios que se obtienen y la forma de obtenerlos, especialmente los referidos a movilización, dieta y ejercicios respiratorios. Si se va a realizar un estoma, la visita a una consulta especializada antes de la intervención mejora mucho los resultados[24-28].

1. Los pacientes deben recibir información oral y escrita completa de lo que se le solicita para mejorar su recuperación después de la cirugía.

Recomendación fuerte +. Nivel de Evidencia moderado.

- Maessen J, Dejong CH, Hausel J, et al. A protocolis not enough to implement an enhanced recovery programme for colorectal resection. Brj Surg 2007;94:224—31[29].

(Otros estudios relevantes[30])

VALORACIÓN DEL RIESGO ANESTÉSICO-QUIRÚRGICO

Los pacientes con mayor riesgo quirúrgico son aquellos de edad avanzada con enfermedad coexistente a los que se les va a realizar un procedimiento de cirugía mayor; estos pacientes tienen una mortalidad postoperatoria de 5-25%, deben ser identificados en la etapa preoperatoria y, tratados de forma óptima antes de la cirugía. Para ello se debe evaluar el riesgo anestésico (mediante escala ASA), el riesgo cardiológico, el quirúrgico y el nutricional; así como, hacer énfasis en los cambios de hábitos nocivos[31,32].

Un paciente informado y optimizado para la cirugía tiene una recuperación más rápida.

EVALUACIÓN DEL RIESGO CARDIOLÓGICO

Según las Guías de Práctica Clínica de la AHA (American Heart Association), los grupos de pacientes con enfermedad cardiológica activa deben ser evaluados y tratados por cardiólogos antes de la intervención quirúrgica[33]:

- **Síndrome coronario inestable**: Angina inestable o Infarto de miocardio reciente.
- **Insuficiencia cardíaca descompensada**: Pacientes con clase funcional NYHA IV o Insuficiencia cardíaca de nueva aparición.
- **Arritmia significativa**: Bloqueo Auriculoventricular Mobitz II o de tercer grado. Arritmias ventriculares o supraventriculares (incluida fibrilación auricular) con respuesta ventricular no controlada (frecuencia cardíaca mayor a 100 latidos/minuto). Bradicardia sintomática.
- **Enfermedad valvular grave**: Estenosis aórtica grave, estenosis mitral grave.

2. Los pacientes con patología cardíaca activa de reciente aparición o descompensada deben ser evaluados por cardiólogos previamente a la intervención.

Recomendación fuerte +. Nivel de Evidencia Alto.

- Fleisher LA, Beckman JA, Brown KA, Calkins H, Chaikof E, Fleischmann KE, et al. ACC/ AHA 2007 Guidelines on Perioperative Cardiovascular Evaluation and Care for Noncardiac Surgery: Executive Summary: A Report of the American College of Cardiology/American Heart Association Task Force on Practice Guidelines (Writing Committee to Revise the 2002 Guidelines on Perioperative Cardiovascular Evaluation for Noncardiac Surgery): Developed in

24

(a) Formato PDF.

(b) Formato TXT.

Figura 1.1: Comparación de una misma página en formato PDF y en formato texto plano.

se enfoque en producir los mejores resultados posibles para este tipo de documento y otros documentos médicos de estructura similar.

La herramienta creada puede tener múltiples casos de uso, en la medida en que convierte en reeditables documentos con un formato inicialmente orientado a la presentación e impresión, y no a la edición. Pero el resultado de procesar documentos PDF a texto estructurado tiene interesantes aplicaciones en el campo del Procesamiento del Lenguaje Natural (PLN). Dentro de los posibles usos de las tecnologías de PLN, con frecuencia es necesario analizar textos y documentos de cierta complejidad. En estos textos, un análisis básico de lenguaje natural con algoritmos como bag-of-words o word2vec puede ser suficiente en problemas como el de "análisis de sentimientos", pero son claramente insuficientes en otras soluciones como los sistemas de "pregunta-respuesta". Es frecuente que, en un texto complejo, un párrafo no haga referencia explícita al tema que trata, sino que lo haga usando referencias a párrafos anteriores o a títulos de capítulo o sección. En estos casos, es imprescindible contextualizar cada párrafo dentro de la estructura jerarquizada del documento completo. Es aquí donde una solución como la desarrollada, que a partir de

documentos PDF extrae el texto junto con su estructura, resulta imprescindible si los documentos a analizar están disponibles exclusivamente en formato PDF, como ocurre con la mayoría de las Guías de Práctica Clínica publicadas.

Como ejemplo de lo expuesto, se ha desarrollado una prueba de concepto de un posible caso de uso, en el que, utilizando Guías de Práctica Clínica convertidas a JSON, se alimenta un chatbot para que sea capaz de responder preguntas sobre diversos problemas de salud.

Capítulo 2

Comparación de librerías

Antes de proceder al diseño y desarrollo del pipeline de procesado, se ha elaborado un estudio en el que se comparan un conjunto de librerías preexistentes que permiten la extracción de texto a partir de documentos PDF, con el objetivo de conocer cuáles son las opciones disponibles y cual sería la solución óptima para tomar como punto de partida.

En esta comparación tiene especial importancia conocer qué librerías proporcionan información adicional sobre el texto más allá de una simple extracción en texto plano, de forma que la posterior estructuración sea lo más sencilla posible. Las librerías estudiadas son: PDFMiner, PDFPlumber, Slate, PDFlib TET y PyPDF2. También se han estudiado dos herramientas de línea de comandos: soffice y pdftohtml.

2.1. PDFMiner

PDFMiner [2] es una librería open source disponible para su uso en Python. Existen dos versiones, una original y descontinuada, y otra mantenida por la comunidad bajo el nombre de pdfminer.six [3].

De cara a la obtención de información adicional sobre el texto, como podría ser fuente y tamaño de letra, en el caso de PDFMiner estos datos pueden ser extraídos dependiendo del formato de extracción en bruto. Cada uno de estos formatos tiene sus ventajas e inconvenientes, que se detallan en los siguientes subapartados. Para probar los diferentes formatos de extracción que ofrece esta librería se ha utilizado un mismo documento PDF, sobre el que además se ha anotado el tamaño del fichero resultante, de cara a conocer cómo de ligero o pesado es cada formato.

2.1.1. Extracción en formato Texto

Es el formato de extracción más simple. Permite extraer texto literal línea por línea sin ofrecer información adicional.

El tamaño del fichero resultante es 5.6 KB.

2.1.2. Extracción en formato Tag Extraction

Ofrece una salida en formato HTML en la que enriquece ligeramente el texto frente a la extracción de texto normal. Realiza las siguientes acciones adicionales frente al formato de extracción anterior:

- Etiqueta el texto por páginas, información que ya se podía conocer en la extracción de texto plano simplemente extrayendo página a página.

- Informa de las coordenadas de inicio del texto de cada página, algo que no es de mucha utilidad si no se acompaña de información adicional sobre el resto de elementos de la página.

- Etiqueta de forma especial algunas palabras que se encuentran en un formato diferente al convencional, pero lo hace de forma errática, etiquetando palabra por palabra, siendo algunas de ellas no relevantes realmente, y dejando sin etiquetar palabras contiguas relevantes.

El tamaño del fichero resultante es 6.3 KB.

Se puede concluir que este formato de extracción no ofrece ninguna ventaja con respecto a la extracción de texto.

2.1.3. Extracción en formato HTML

Este formato de extracción supone una mejora sustancial en lo que a enriquecimiento del texto se refiere respecto a los dos mencionados anteriormente. A continuación se describe la serie de cambios que realiza en la extracción:

- Agrupa texto en pequeños párrafos de acuerdo a sus coordenadas de inicio, fuente y tamaño de letra.

- En algún caso separa el texto más de lo debido, pero es posible que de acuerdo a datos como la fuente y el tamaño de letra se pueda realizar una reconstrucción más apropiada de la separación entre bloques de texto.

- Indica algunos datos ajenos al texto como bordes que lo rodean, líneas y cajas, que realmente no son útiles de cara a un posterior procesamiento del lenguaje natural.

El tamaño del fichero resultante es 21.5 KB.

Estas cualidades hacen que posiblemente sea el mejor formato de extracción, teniendo en cuenta el equilibrio entre información adicional obtenida y tamaño del archivo, que posteriormente se traduce en tiempo de procesamiento.

2.1.4. Extracción en formato XML

Es el formato de extracción en bruto que más información ofrece:

- Indica la posición de cada uno de los elementos que extrae, aunque no se trate de texto.
- Además, indica para cada carácter su posición, fuente y tamaño de letra.

El tamaño del fichero resultante es 586 KB.

Pese al alto nivel de detalle en la información que provee, este mismo factor no resulta viable de cara a un procesamiento rápido del texto, siendo probablemente mucho más lento y complejo que en formato HTML.

2.2. PDFPlumber

PDFPlumber [4] es una librería de extracción de texto de PDFs construida sobre PDFMiner. Dispone de distintos modos de extracción con los que obtener información carácter a carácter (fuente, tamaño de letra, posición...). También permite la extracción de tablas existentes en el PDF.

La información útil para la posterior reconstrucción de la estructura del texto no difiere de la que se puede obtener con PDFMiner en su modo de extracción a XML (aunque en ese caso sea necesario un parseo posterior). Para obtener información de fuente y tamaño de letra (lo más relevante además del texto que ofrece una librería), es necesaria la inspección carácter a carácter. Es por esto que para ese fin, al igual que se ha concluído en la extracción de PDFMiner en formato XML, la extracción de dicha librería a formato HTML es una opción más adecuada.

2.3. Slate

Al igual que PDFPlumber, Slate [5] es un paquete de Python construido sobre PDFMiner. Su punto fuerte es la sencillez con la que permite la extracción básica de texto. Sin embargo, tal y como se anuncia en su documentación, si se desea obtener información detallada sobre el texto, como fuente o tamaño de la letra, es necesario

hacerlo a través de la API de PDFMiner. Por tanto, esta herramienta no supone ninguna ventaja respecto a su librería base.

2.4. PDFlib TET

PDFlib TET (PDFlib Text and Image Extraction Toolkit) [6], es un software comercial que permite la extracción de texto de un PDF. Las ventajas que su documentación dice tener son claras:

- Extracción básica de texto refinada: texto sombreado, texto con acentos, unión automática de palabras separadas entre líneas por un guión, etc.
- Información valiosa sobre el texto: fuente, tamaño de letra, color, etc.
- Separación de párrafos en base a sus encabezados: precisamente uno de los objetivos clave que se desea lograr, para así tener una estructuración del texto más acorde a la intención informativa del documento.
- Eliminación de algunos encabezados y pies de página: Los encabezados y pies de página que se encuentran marcados con el tag Artifact son eliminados del resultado final de extracción.
- Soporte para múltiples lenguajes de programación: El software se encuentra disponible para multitud de lenguajes de programación, como podrían ser C, C++, Java, Python, Ruby, .NET, Swift...
- Programado en C y C++: El hecho de que la librería haya sido programada en estos lenguajes hace que la ejecución de sus funciones internas sea mucho más rápida que la que tendría la misma librería desarrollada en un lenguaje interpretado como Python, o incluso en otros compilados como Java.

Por otro lado, tiene una desventaja que hace que sea casi inmediata la decisión de desechar la posibilidad de utilizar esta herramienta. Es precisamente que se trata de un software comercial, cuya licencia tiene un alto coste económico. En su versión de escritorio, disponible para Windows 7, 8 y 10 y MacOS (no está disponible para Linux), el precio es de 415€. Por otro lado, el coste de la licencia para servidor, necesaria para poder utilizar la herramienta en Linux (también disponible para Windows Server), alcanza 1095€[1] .

[1]Ambos precios, a la fecha de realización de este trabajo, en junio de 2020

PDFlib TET posee una versión de prueba gratuita disponible para documentos de hasta 10 páginas, con la que a partir de unos simples comandos se puede extraer texto de un PDF al formato TETML (TET Markup Language), que tiene similitudes con XML. Se ha podido comprobar que, en su modo de extracción `page` [7], ofrece resultados muy buenos a la hora de realizar tareas como la fusión de líneas pertenecientes a un mismo párrafo. Pese a esto, se puede observar que, por ejemplo, no consigue manejar bien la unión de párrafos separados por un salto de página debido a los números de página. Tampoco provee de forma directa una jerarquía de párrafos que se pueda utilizar para la reconstrucción de la estructura original del documento.

Más allá del coste de la licencia, basar el trabajo a desarrollar en un software comercial impide su posterior distribución como herramienta open source, de forma que pudiera no sólo ser utilizada en una ocasión para el análisis de textos clínicos, sino también en cualquier dominio de aplicación en el que pudiera ser útil.

2.5. PyPDF2

En cuanto a lo que a extracción simple se refiere, PyPDF2 [8] consigue extraer el texto del PDF en su totalidad o práctica totalidad. Sin embargo, tiende a juntar palabras (eliminar espacios en blanco necesarios) en más ocasiones que PDFMiner y sus librerías derivadas.

2.6. Línea de comandos

Otra alternativa para la extracción de texto de un PDF es la utilización de distintos comandos que incorporan los sistemas Unix. De esta manera es posible lograr una extracción de un PDF a un formato de texto, como por ejemplo HTML. Esto se puede lograr gracias al paquete **soffice** [9] simplemente utilizando el siguiente comando:

```
soffice --convert-to html ./input_document.pdf
```

Sin embargo, la extracción a HTML que ofrece no se diferencia en cuanto a información sobre el estilo del texto de una extracción básica.

Otro paquete que permite una extracción enriquecida de un documento PDF mediante línea de comandos es **pdftohtml** [10]. Con el siguiente comando se puede conseguir una extracción a HTML:

```
pdftohtml -enc UTF-8 -i -s -xml -noframes input_doc.pdf output.html
```

Este paquete presenta resultados interesantes, como obtener la fuente de letra al igual que PDFMiner, pero con un fichero de salida con formato más comprimido. La desventaja es que no proporciona información sobre el tamaño de la letra, algo que sí

que extraen otras librerías. Otra desventaja es la extracción de texto en páginas con texto rotado (orientación de 90 o 270 grados), que provocan que la sintaxis de salida del HTML se rompa.

2.7. Conclusiones de la comparación

Tras realizar esta comparación entre librerías, se ha llegado a la conclusión de que el mejor punto de partida, en lo que a herramientas open source y freeware se refiere es PDFMiner, haciendo uso de su modo de extracción de PDF a HTML. De este modo, se obtiene información de posición del texto, tamaño y tipo de fuente, con un uso de memoria mucho menor que si se utilizara el formato XML.

Capítulo 3

Componentes del pipeline

En este capítulo se describen cada uno de los componentes que posee el pipeline de procesamiento desarrollado.

3.1. Esquema general

En el esquema de la Figura 3.1 se muestran de forma conceptual los pasos de procesamiento que son llevados a cabo por el programa, los cuales se detallan en los siguientes apartados de este capítulo.

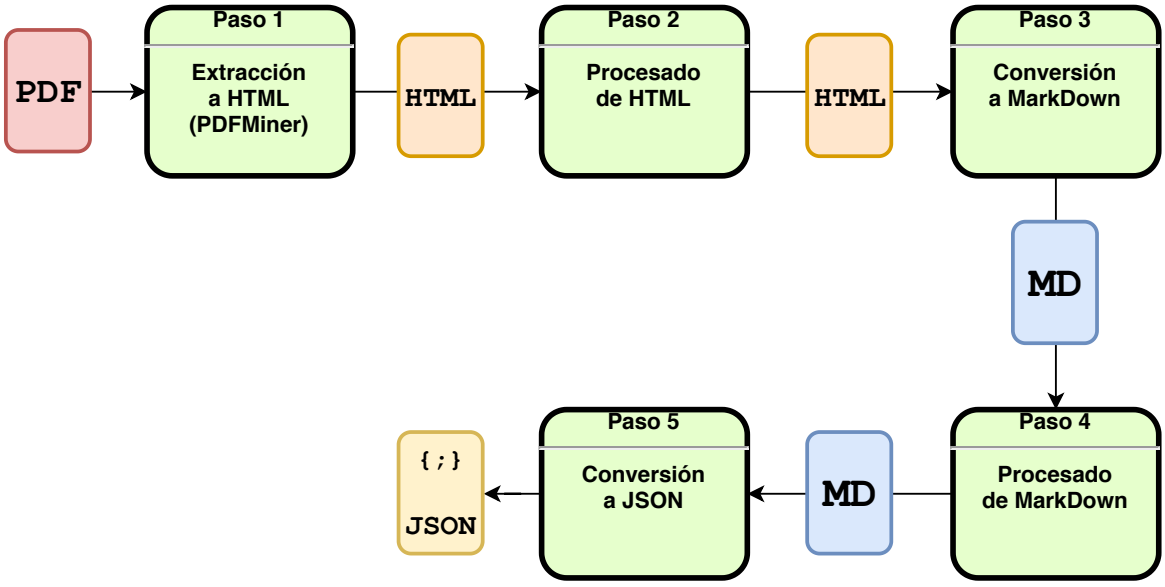


Figura 3.1: Esquema general del pipeline de procesamiento.

3.2. Extracción a HTML

El primer paso del procesamiento es la extracción del texto del PDF. Para esto se ha utilizado la librería PDFMiner, en su versión pdfminer.six. Hacer uso de esta

herramienta ha implicado utilizar Python como lenguaje de programación para el desarrollo del pipeline, aunque esta decisión no solo se debe a la utilización de PDFminer, sino también para lograr un desarrollo rápido con un número de líneas de código reducido.

La extracción se ha realizado activando el flag `detect_vertical` en el constructor del intérprete de PDFMiner. Esto facilita la extracción de texto cuya orientación no es la estándar, sino que se encuentra rotado 90 o 270 grados.

La existencia de texto rotado es un problema, ya que la extracción directa que ofrece PDFMiner para este tipo de texto suele resultar en que cada uno de los caracteres en cuestión esté separado por un salto de línea en el texto extraído.

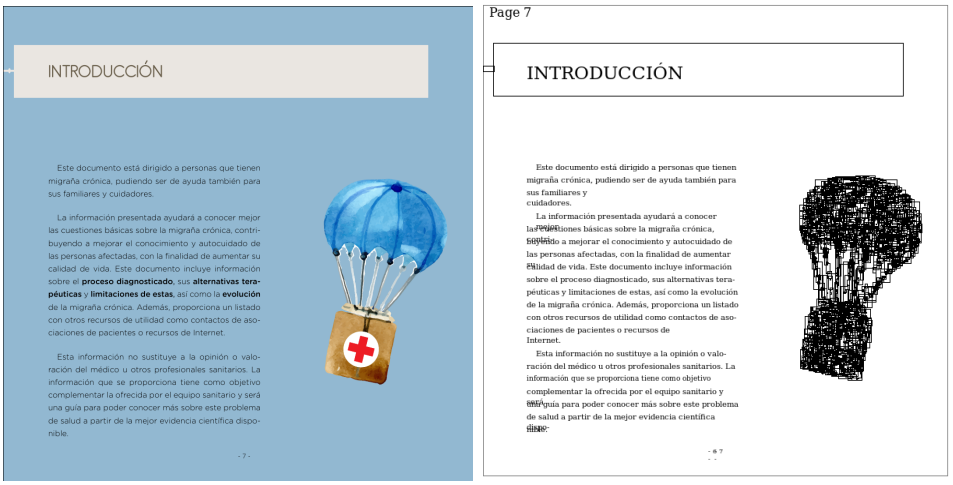
En las Guías de Práctica Clínica, los fragmentos de texto rotado no suelen ser relevantes. Principalmente se tratan de un mismo texto que se repite en el lateral de todas las páginas del documento o páginas enteras rotadas (aunque la orientación de la página en el PDF sea vertical). Por suerte, el texto en los laterales debe ser eliminado y las páginas cuyo texto completo está rotado suelen tratarse de tablas presentes en anexos que no aportan demasiada relevancia al documento, por lo que este tipo de texto puede ser eliminado en siguientes fases del pipeline.

De todas maneras, en el caso de que se desee extraer también este tipo de texto, se ha diseñado un sistema para reconocer las páginas cuya mayor parte del texto está rotado y así poder extraer el texto de nuevo habiendo posicionado la página en horizontal. Este sistema aprovecha el hecho de que procesar una misma página activando el flag `detect_vertical` y desactivándolo genera prácticamente los mismos resultados en una página cualquiera, pero no en una cuyo texto esté girado 90 o 270 grados. Esto es fácilmente reconocible a través de la diferencia en el número de ocurrencias de la expresión regular `\w\n`, es decir, una letra o número seguido de salto de línea. También se puede reconocer debido a que una página de estas características suele tardar varias veces más en procesarse si no se utiliza el flag `detect_vertical`, de forma que a través de un timeout basado en el tiempo empleado con el flag activado la detección también es efectiva. En el caso de que se detecte que gran parte de la página contiene texto rotado, la página se vuelve a extraer tras posicionarla en horizontal, girándola en sentido horario, dado que en todos los documentos probados, el texto se encuentra rotado en sentido antihorario en este tipo de páginas.

Tras probar ambas modalidades de extracción, directa y utilizando la técnica para reconocer páginas con texto rotado, se ha llegado a la conclusión de que para las Guías de Práctica Clínica es mucho más productivo realizar una extracción directa, ya que el otro modelo emplea al menos el doble de tiempo (hay que realizar mínimo 2 procesados por página) para obtener una información adicional reducida y poco relevante en la

mayoría de casos.

El resultado en cualquier caso es una conversión de PDF a HTML, de forma que se dispone, además del texto existente en el documento, de información sobre su posición, tamaño de letra y fuente.



(a) Formato PDF. (b) Formato HTML.

Figura 3.2: Comparación de una misma página antes y después de la extracción.

3.3. Procesado de HTML

3.3.1. Eliminación de elementos gráficos

La eliminación de elementos gráficos supone preservar en formato HTML únicamente los contenedores `div` de texto. Esto facilita y agiliza los posteriores pasos del procesamiento, ya que el tamaño que ocupa en memoria es menor. La eliminación se hace por medio de expresiones regulares, detectando los contenedores cuya estructura interna sea propia de texto, y anexionándolos a una nueva cadena de texto, que se utilizará por el siguiente componente del pipeline como punto de partida.

3.3.2. Eliminación de encabezados y pies de página

Uno de los puntos claves del pipeline de procesamiento es precisamente la eliminación de encabezados y pies de página. Este concepto se refiere a frases repetidas en todas las páginas del documento, como podría ser el título de la GPC (Guía de Práctica Clínica), el nombre del apartado o capítulo, o mismamente los números de página. Este texto, debido a su posición en los extremos inferiores o superiores de una página, en una extracción directa cortaría los párrafos que empiezan en una página y continúan en la siguiente, impidiendo una correcta reconstrucción del contenido.

Para esto se ha elaborado una técnica que reconoce cuáles son los límites en píxeles (medida utilizada en el HTML extraído para indicar posición) para lo que se puede entender como contenido relevante, de manera que se adapte a páginas de distintos tamaños y orientaciones. Cabe mencionar que esta primera detección de límites se realiza de forma previa a la eliminación de elementos gráficos, mientras que el paso en el que sí se modifica el HTML descartando encabezados y pies de página se realiza después.

En el resultado de extracción a HTML que realiza PDFMiner, resultan relevantes los tipos de líneas mostrados en la Figura 3.3.

```
<span style="position:absolute; border: gray 1px solid; left:0px; top:50px; width:595px;
height:841px;"></span>
<div style="position:absolute; top:50px;"><a name="1">Page 1</a></div>
```

```
<span style="position:absolute; border: gray 1px solid; left:0px; top:8077px; width:595px;
height:841px;"></span>
<div style="position:absolute; top:8077px;"><a name="10">Page 10</a></div>
```

Figura 3.3: Fragmentos HTML donde se indica la posición de inicio de página.

Estas líneas indican para cada página la altura en píxeles (teniendo en cuenta que vale 0 al principio del documento y alcanza su valor máximo al final) donde comienza la página o `top`, y la altura de la propia página o `height`. De esta forma, según indica este formato, la página `i` tiene unos límites marcados por:

$$(top[i], top[i] + height[i]) \tag{3.1}$$

El número de página que indican estas líneas puede ser erróneo, pero el orden en el que aparecen sí que se corresponde con las páginas reales del documento, por lo que los datos de posición pueden ser utilizados si se asocian con una página por su orden de aparición.

Teniendo en cuenta estos límites como los límites donde puede aparecer texto en una página, se quieren generar unos nuevos límites, que suponen una reducción del rango, para desechar todo el texto que no se encuentre dentro de ellos, asumiendo que se trata de encabezados o pies de página.

Para esto se han establecido unos coeficientes, `kl` y `ku`, de forma que los nuevos límites inferior (parte superior de la página) y superior (parte inferior de la página) quedan de la siguiente forma:

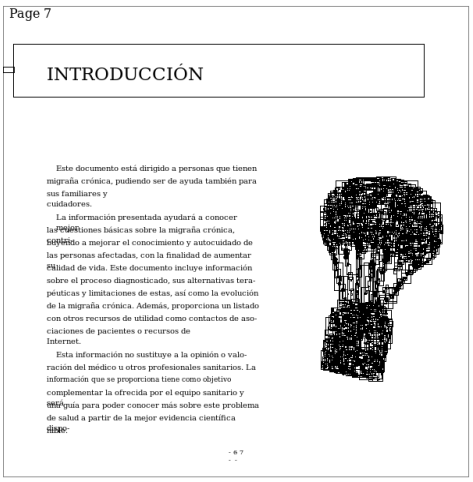
$$lower\_bound = top + kl * height \tag{3.2}$$

$$upper\_bound = (top + height) - ku * height \tag{3.3}$$

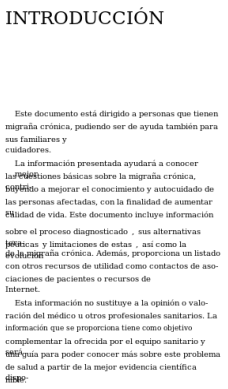
Una vez se tiene esta información, se parsean los bloques de texto del fichero, obteniendo uno por uno su posición en píxeles en el documento. Si alguno de ellos no se encuentra dentro de los límites aceptados, se descarta.

Tras probar múltiples coeficientes, se ha encontrado que los que arrojan mejores resultados, es decir, un mayor equilibrio entre eliminación de encabezados y preservación de contenido relevante, son 0.05 para `kl` y 0.08 para `ku`.

A la hora de descartar, también se comprueba el tamaño de fuente del bloque de texto analizado, ya que si es mayor que un umbral, se entiende que el texto es un título, de forma que nunca será descartado, independientemente de su posición. El umbral escogido para el tamaño de letra, debido a sus buenos resultados en experimentación con el conjunto de documentos de prueba, es 18.



(a) Formato HTML.



(b) Formato HTML depurado.

Figura 3.4: Comparación de una misma página antes y después de eliminar elementos gráficos y pies de página.

3.3.3. Eliminación de texto vertical o rotado

Este componente del pipeline, a través de una expresión regular, consigue identificar texto vertical, que en la mayoría de los casos se corresponde con la extracción de texto rotado en el documento PDF. Este texto, debido a sus características no permite una correcta reconstrucción del contenido, así que sumado a que no suele ser relevante (en muchas ocasiones se trata del nombre del apartado o el título del documento repetido

a lo largo de todas las páginas en un lateral), se ha procedido a detectarlo y eliminarlo. La clave para construir la expresión regular que lo detecta es que se cumple un patrón consistente en una secuencia de varios bloques de texto consecutivos de muy pocas letras (entre 1 y 5).

3.3.4. Ordenación de contenedores de texto

El paso final dentro del procesamiento de HTML consiste en ordenar los contenedores de texto de acuerdo a su posición en el documento. Si se visualiza el HTML procesado hasta este momento se puede apreciar que los elementos están en su posición correcta, debido a que ésta se indica en píxeles para cada uno de los contenedores. Sin embargo, los contenedores no siempre se encuentran ordenados en el documento, es decir, en algunos casos se ha extraído antes un contenedor que se posiciona en la parte inferior de una página, que uno que se posiciona en la parte superior.

Por esta razón, se ha creado una lista siendo cada uno de sus elementos una tupla compuesta por el texto HTML y la métrica `top` en píxeles de dicho contenedor. Posteriormente los elementos se han ordenado de acuerdo a un orden ascendente de `top`. Como método de ordenación se ha escogido el ordenamiento por inserción, de forma que se itera sobre la lista y, cuando se encuentra un elemento con un `top` menor que el anterior, se recoloca mirando uno a uno los elementos previos hasta que está bien posicionado.

El método descrito en el anterior párrafo tiene coste máximo cuadrático $O(n^2)$ sobre el número de elementos en la ordenación de una lista cualquiera; sin embargo, este coste nunca se da en la práctica. La extracción a HTML proporcionada por PDFMiner en efecto desordena algunos elementos, pero la gran mayoría de ellos están ordenados, y para los que no lo están, se encuentra su posición correcta tras unas pocas comprobaciones de elementos anteriores, ya que en ningún caso existe más de una página de diferencia (el número de elementos en una página dependerá del documento) entre la posición inicial y la final una vez se ha ordenado.

Métrica por documento	Media
Bloques totales	3496.85
Bloques desordenados	972.08
Proporción desordenados	0.28
Distancia media posición inicial-final	12.78

Tabla 3.1: Estadísticas de ordenación de bloques en un conjunto de 40 documentos.

3.4. Conversión a MarkDown

Después de haber realizado un procesado sobre el documento en HTML, el siguiente paso es la transformación a MarkDown. Se ha escogido este formato debido a que se asemeja en gran medida al texto plano, de forma que realizar una depuración sobre él utilizando expresiones regulares es mucho más simple que, por ejemplo, sobre HTML o JSON, y permite además incluir información sobre títulos (con varios niveles de relevancia) a través de iniciar una línea con una secuencia de caracteres `#`.

La conversión a MarkDown toma los bloques de texto en HTML y los añade a una nueva cadena de texto en la cual se forma el documento en el nuevo formato. La modalidad en la que se añade cada bloque de texto (como texto estándar o como título de nivel X) depende del tamaño de letra que tenga ese bloque. Como cada documento es diferente, se realiza un análisis previo de las fuentes de letra existentes en el documento y su frecuencia, el cual se describe a continuación.

3.4.1. Análisis de tamaños de fuente

El análisis de tamaños de fuente comienza por conocer el número de apariciones (número de caracteres) de cada tamaño. Para ello se identifican los bloques de texto en HTML mediante una expresión regular, y se itera sobre cada uno de ellos para cuantificar la longitud en caracteres acumulada del texto literal en base al tamaño de fuente del bloque. Un ejemplo del análisis de uno de los documentos se puede visualizar en el histograma de la Figura 3.5. Una vez se dispone de esta información se pueden obtener diversos umbrales útiles para la conversión a MarkDown, tal y como se explica en los siguientes subapartados.

Cabe destacar que el análisis forma parte del pipeline de procesamiento, por lo tanto los umbrales varían y se adaptan dependiendo de cada documento.

Umbral para texto estándar

El primer objetivo del análisis de tamaños de fuente del documento es determinar a partir de qué tamaño un texto deja de ser texto estándar y se considera título.

Para lograr esto, en una primera versión se optó por seleccionar el tamaño más común del documento, tomando como referencia el número de caracteres que tienen asignado dicho tamaño. Esta aproximación es lo suficientemente buena en algunos documentos, pero en otros en los que hay dos o más tamaños altamente frecuentes, no funciona como cabría esperar.

En la segunda versión, en lugar de utilizar el tamaño de fuente más común, se ha pasado a escoger el umbral de forma que ese tamaño y todos los inferiores, alcancen

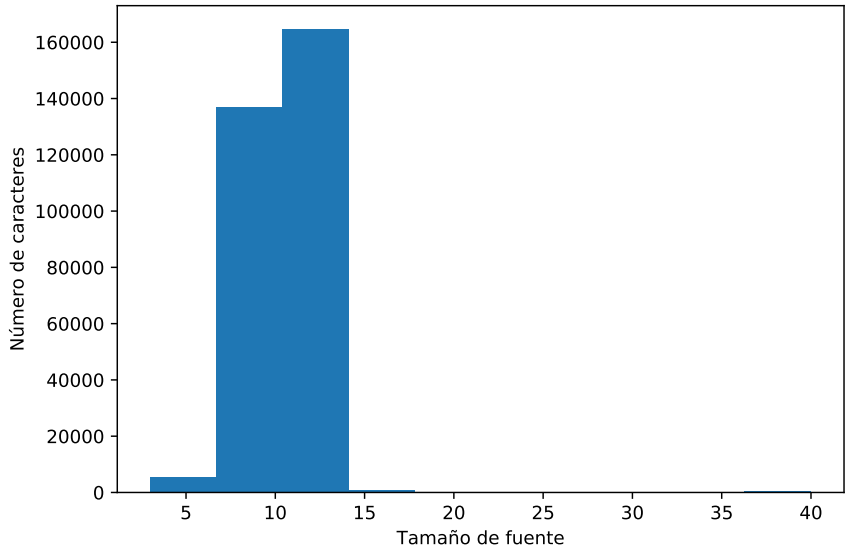


Figura 3.5: Histograma de tamaños de fuente de un documento.

entre la suma de sus caracteres el 95 % del total de caracteres del documento. En algunos casos coincide con el tamaño más común, y en otros casos el umbral aumenta, obteniendo en tales casos mejores resultados.

Otra mejora realizada es obviar del cálculo todos aquellos caracteres cuyo tamaño de fuente sea 0. En algunos documentos, los espacios en blanco son extraídos como texto de tamaño 0, constituyendo un porcentaje más que considerable del total de caracteres, haciendo que los cálculos basados en porcentajes no se ajusten a los que se obtendrían idealmente según el algoritmo diseñado.

Umbral para texto de referencias bibliográficas

Otro umbral que es interesante determinar es el tamaño máximo para que un número sea considerado una referencia a bibliografía. Un ejemplo de esta situación se puede ver después de la palabra "duración" en el texto de la Figura 3.6. Este umbral puede ser utilizado posteriormente para eliminar estas referencias del texto, ya que no forman parte del contenido como tal y pueden ser confundidas con otros números y texto realmente relevante.

de a experiencias semejantes a las psicóticas, pero que difieren de los síntomas psicóticos francos en su intensidad, frecuencia y/o duración.[52] Como es lógico suponer, no estarán

Figura 3.6: Referencia bibliográfica en uno de los documentos de prueba.

En el análisis, se ha determinado cuál es el tamaño de fuente más grande que se encuentra dentro del intervalo del 10 % de caracteres más pequeños del documento.

También se ha identificado el número correspondiente al 80 % del umbral para tamaño estándar. Finalmente, el tamaño elegido para ser tomado como umbral para referencias bibliográficas es el máximo entre los dos mencionados en este párrafo.

Jerarquía de títulos

Como se ha establecido anteriormente, todo el texto que queda por encima del tamaño de fuente umbral, ha sido considerado título. En cualquier documento de esta naturaleza, existe una jerarquía de títulos, que el lector reconoce ya sea un por un número identificador, el tamaño de la fuente empleada u otros factores.

Para hacer una distinción automática y aproximada de estos niveles de título, se ha utilizado como factor discriminante el tamaño de la fuente, que también se ha usado para determinar qué caracteres formaban parte del "cuerpo de texto" o texto estándar. El objetivo es agrupar los distintos tamaños de fuente "grande" existentes en el documento en varios intervalos, de forma que se asocie cada uno de esos intervalos con un nivel de título. El número de niveles diferentes máximo es 6, ya que es el máximo que permite MarkDown, aunque sin ningún problema se podría haber utilizado otro, debido a que este formato no es el final, sino JSON, y no hay por qué adherirse al 100 % al estándar de MarkDown.

En una primera versión para la selección de los intervalos se ha hecho uso del algoritmo de clustering K-medias [11], partiendo de un conjunto de datos consistente en los tamaños de fuente existentes en el documento mayores que el umbral.

En la versión utilizada finalmente, se ha optado por emplear otro tipo de método, ya que K-medias está pensado para datos multidimensionales, además de ofrecer una solución aproximada, no la óptima, que en base a aumentar el número de repeticiones del algoritmo hay más probabilidades de encontrar. Se ha hecho uso del algoritmo de clustering para datos unidimiensionales Ckmeans.1d.dp [12], propuesto por Haizhou Wang y Mingzhou Song, que, mediante programación dinámica, provee la solución óptima al problema de K-medias en tiempo $O(n^2k)$. Realmente la complejidad temporal no es demasiado relevante ya que los datos con los que se trabaja y se prevé trabajar son siempre de un tamaño muy pequeño (rara vez hay más de 20). Utilizar este algoritmo permite ofrecer un resultado determinista frente a un mismo conjunto de datos. El número de clusters elegido es el valor mínimo entre 6 (número de niveles de título diferentes que ofrece MarkDown) y el número de datos existentes. En la mayoría de casos hay más de 6, por lo que éste es finalmente el número de clusters en la mayor parte de las situaciones. Este algoritmo y su implementación [13], devuelve directamente los intervalos dentro de los que se encuentra cada uno de los tamaños de fuente de entrada.

Tras la ejecución del algoritmo, se ha designado a cada uno de los intervalos una

jerarquía de título, haciendo que los datos (tamaños de fuente) que tienen asignado el intervalo con valores más altos, posean el nivel 1 de título, los que tengan asignado el intervalo con los segundos valores más altos, posean el nivel 2, y así sucesivamente. El resultado es un mapa clave-valor (diccionario en Python) donde las claves son cada uno de los tamaños de fuente mayor que el umbral de texto estándar y los valores el nivel de título que cada uno de ellos tiene asignado.

3.4.2. Conversión

Una vez realizado el análisis de tamaños de fuente del documento, se conoce cuál es el umbral fijado para el texto estándar, cuál es el umbral para reconocer números de referencias bibliográficas y se ha obtenido el diccionario que relaciona cada uno de los tamaños de fuente mayores que el umbral de texto estándar con los niveles de título. Con esta información, se sigue un algoritmo para la conversión del HTML a MD que se puede estructurar en una serie de pasos.

Mediante una expresión regular se identifican cada uno de bloques de texto del HTML. Se obtiene una lista con todas las ocurrencias y para cada uno de los elementos se hace lo siguiente:

1. Se obtiene el texto literal que contienen y el tamaño de fuente.

2. En el texto literal, se reemplazan todos los caracteres salto de línea por `<br>`. Esto evita problemas sobre todo en la creación de títulos en el texto MarkDown resultante.

3. En el texto literal, se reemplazan todos los caracteres `#` precedidos por inicio de cadena o `<br>` por `\#`. Esto permite diferenciarlos posteriormente de líneas que realmente se tratan de título en MarkDown debido a su tamaño de fuente original.

4. Si el texto literal se trata de una secuencia de dígitos, o de dos o más números separados por comas o guiones, y el tamaño de fuente es menor o igual que el umbral para referencias bibliográficas, se descarta. También se descarta si el tamaño de fuente es menor que 1 (en algunos documentos los espacios en blancos son extraídos como texto con tamaño 0).

5. Si el tamaño del texto en el anterior elemento de la lista (en el caso de procesar primer elemento se escoge 0 como valor) y el tamaño actual son menores o iguales que el umbral para texto estándar, se añade a la cadena resultante un salto de

línea seguido del texto literal del elemento. Esto significa que el texto se ha identificado como estándar.

6. Si no se cumple la condición anterior y el tamaño actual es el mismo que en el elemento anterior, a la cadena resultante se le añade un espacio en blanco seguido del texto del elemento. Esto significa que el texto se ha identificado como continuación de un título (es necesario que continúe en la misma línea).

7. Si ninguna de las condiciones anteriores se cumple, y el tamaño de fuente actual es mayor que el umbral para el texto estándar, el elemento en cuestión se corresponde con el inicio de un título. A la cadena resultante se le añade un salto de línea, seguido de tantos caracteres **#** como indique el nivel de título asociado al tamaño de fuente, un espacio en blanco y finalmente el texto literal.

8. Si ninguna de las condiciones anteriores se cumple, se trata de texto estándar, por lo que se añade a la cadena resultante un salto de línea seguido del texto literal.

El resultado es una cadena texto en MarkDown, la cual va a ser procesada a su vez en distintos pasos, los cuales se describen en el siguiente apartado.

(a) Formato HTML depurado. (b) Formato MD.

Figura 3.7: Comparación de una misma página antes y después de convertir de HTML a MarkDown.

3.5. Procesado de MarkDown

El procesamiento del documento transformado a MarkDown se compone de diversos pasos de depuración, con el objetivo de realizar una separación de párrafos, conteniendo

cada uno de ellos en una línea, además de arreglar diversos fallos en el texto generados por el extractor de PDFMiner.

Este componente del pipeline se compone de un gran número de pasos de procesamiento que realizan una pequeña acción sobre el texto que han depurado los anteriores, y que se describen a continuación.

Reemplazar por salto de línea

Reemplaza `<br>` en el texto por un carácter de salto de línea si la línea es estándar, y lo reemplaza por un espacio en blanco si la línea es un título.

Eliminar falsos títulos

Elimina las líneas de título que no contienen ninguna palabra.

Eliminar líneas vacías

Elimina las líneas en blanco del documento.

Reemplazar caracteres CID

Sustituye a los elementos cid (un formato para codificar algunos caracteres extraídos por PDFMiner) por su equivalente en caracteres UNICODE (no es 100 % exacto, especialmente en caracteres ASCII extendidos).

Reemplazar por guión

Reemplaza un conjunto de símbolos con significado similar al guión por '-' en el texto provisto.

Unir por guión

Une las líneas que están separadas por una palabra dividida por un guión, de forma que la palabra en cuestión se reconstruye.

Unir líneas en párrafos

Procesa el texto de manera que las líneas que forman parte del mismo párrafo semántico se fusionan en una sola línea.

Unir et al.

Une las líneas del texto separadas por el punto en 'et al.' cuando se encuentra al final de una línea.

Unir símbolo beta

Une las líneas del texto separadas por el carácter β cuando se ubica al final de una línea.

Unir vs.

Une las líneas del texto separadas por el punto en 'vs.' cuando se encuentra al final de una línea.

Arreglar símbolo ñ

Reconstruye el carácter `ñ` si está roto tras la extracción (`~n`).

Unir puntos suspensivos

Fusiona las líneas que están separadas por puntos suspensivos situados al final de una línea, cuando se supone que su continuación forma parte del mismo párrafo.

Unir restas

Fusiona las líneas que están separadas debido a una línea que comienza con un guión, cuando en realidad se trata de una resta que comienza en la línea anterior.

Unir por ':'

Fusiona las líneas separadas por dos puntos (:) al final de una línea cuando la siguiente comienza con minúsculas.

Eliminar guiones duplicados

Elimina los guiones consecutivos duplicados del texto.

Arreglar signos de puntuación

Elimina los espacios en blanco innecesarios entre algunas palabras y signos de puntuación adyacentes.

Unir preguntas en títulos

En algunos documentos, un solo título puede estar compuesto por palabras de diferentes tamaños de fuente, por lo que al procesarlo puede separarse en varias líneas con distinto nivel de título. Este procedimiento fusiona las líneas en una sola cuando el título es una pregunta, ya que puede ser fácilmente detectado (al menos en español)

debido a los símbolos '¿' y '?'. El nivel de título resultante en la línea fusionada es el más alto (menos significativo) de los niveles que componían la pregunta.

Eliminar líneas no relevantes

Elimina las líneas que no contienen ninguna palabra ni ningún número.

Eliminar espacios duplicados

Elimina los espacios en blanco duplicados en el texto, y también elimina los espacios en blanco al principio y al final de las líneas.

Eliminar cadenas repetidas

Elimina las subcadenas de texto que son idénticas y consecutivas (repetidas más de 3 veces), que suelen representar un fallo de la extracción.

En la Figura 3.8 se muestra un ejemplo de un fragmento de documento antes y después de los pasos de depuración de MarkDown, donde se puede observar, entre otras cosas, la correcta reconstrucción de los párrafos en una única línea.

(a) Formato MD. (b) Formato MD depurado.

Figura 3.8: Comparación de una misma página antes y después del procesamiento en formato MarkDown.

3.6. Conversión a JSON

Una vez se ha realizado la depuración sobre el texto del documento en MarkDown, el último paso es la conversión a JSON.

El resultado es un fichero cuyo nodo raíz sigue el formato presente en la Figura 3.9.

Figura 3.9: Nodo raíz del documento en formato JSON.

El valor de `content` es una lista que donde cada nodo tiene un nivel inferior que cualquiera de sus hijos, considerando hijos a aquellos subnodos que se encuentren dentro de la lista `content` del nodo padre. Un nodo no tiene por qué tener hijos en absoluto. Un nodo cualquiera (excluyendo al nodo raíz descrito anteriormente) se construye de la forma que se puede ver en la Figura 3.10.

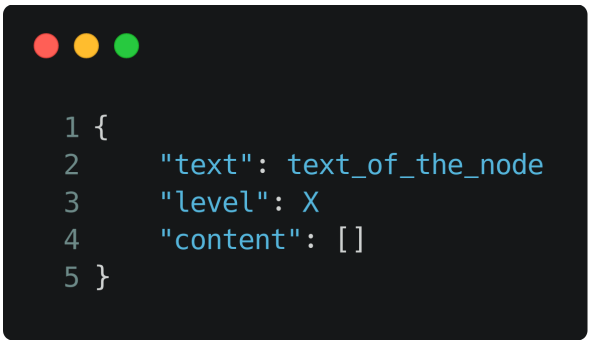


Figura 3.10: Nodo cualquiera del documento en formato JSON.

Los niveles se encuentran en el rango [0, 7], donde 0 se corresponde únicamente el nodo raíz del documento, 7 hace referencia al texto estándar, y del 1 al 6 son los distintos niveles de títulos, siendo el 1 el más significativo y 6 el que menos.

En el Anexo B se encuentra un ejemplo de extracción y procesado de un documento, de manera que se pueden observar con más detalle los resultados finales que se describen en este apartado.

Capítulo 4

Prueba de concepto de un caso de uso: chatbot

El trabajo desarrollado en este Trabajo Fin de Grado tiene como objetivo la extracción de texto estructurado a partir de Guías de Práctica Clínica en formato PDF. Esta herramienta da pie a diversos proyectos que tomen este trabajo como punto de partida.

Una de estas posibles continuaciones es la creación de un chatbot, de manera que un usuario haga una pregunta sobre un aspecto concreto o general de medicina, y el sistema le responda con información que le sea de utilidad.

En este caso de uso la herramienta se utilizaría para obtener el texto estructurado de Guías de Práctica Clínica, y con él entrenar el sistema de forma mucho más sencilla y eficiente que si se tuviese a disposición simplemente el texto plano del documento PDF.

4.1. Versión básica con Dialogflow

En un primer momento se ha comenzado a desarrollar el chatbot utilizando Dialogflow [14], una herramienta de Google que facilita a desarrolladores la creación de chatbots y que internamente utiliza Machine Learning para proveer PLN (Procesamiento del Lenguaje Natural).

Esta primera versión de la prueba de concepto se ha basado en crear un *intent* por Guía de Práctica Clínica, e introducir manualmente las frases de prueba, que se trataban de preguntas presentes de forma literal en títulos de apartados del documento, p. ej. “¿Qué es la depresión?”, “¿Qué causa la depresión?”, etc. Posteriormente se ha probado que al introducir una pregunta sobre un problema de salud sobre alguna de las Guías de Práctica Clínica usada como prueba, el sistema devuelva el *intent* correspondiente al documento.

Un *intent* es lo que permite describir una petición típica por parte del usuario que interactúa con el chatbot. Los *intents* llevan asociadas una o más frases de entrenamiento, que sirven como ejemplo para que, al reconocer una frase igual o parecida, el chatbot sepa cuál es el *intent* al que se está haciendo referencia. Siguiendo el ejemplo anterior, preguntas similares a "¿Qué es la depresión?" o "¿Qué causa la depresión?", hacen que el chatbot detecte que el *intent* al que intenta hacer referencia el usuario es el correspondiente a la Guía de Práctica Clínica sobre la depresión.

Tras comprobar gracias a Dialogflow, que realizar la prueba de concepto de chatbot es viable, se ha procedido a realizar un programa que permite introducir frases de entrenamiento para *intents* de forma masiva, sin tener que pasar por el proceso manual de inserción una a una.

Una de las opciones era seguir utilizando Dialogflow, creando un script en Python que a través de la API de Google Cloud introdujera en el sistema las frases de entrenamiento.

Sin embargo, se ha optado por buscar otra alternativa open source, que permita hacer un entrenamiento del modelo de Machine Learning offline, sin tener que enviar datos en ningún momento a terceros, algo que es más prudente en el sector de la medicina, aun cuando las Guías de Práctica Clínica son de dominio público. Otra de las razones para elegir una herramienta de código abierto es, como se ha comentado más arriba, la no dependencia de productos propietarios de cara a una reutilización del producto resultante como herramienta de código abierto. Por último, la posibilidad de alimentar y parametrizar las distintas herramientas de chatbot mediante APIs públicas, hace que la migración de una herramienta a otra sea bastante fácil, en el caso de que un cliente decidiera en última instancia optar por una solución propietaria. La opción elegida finalmente, que cumple con estas condiciones es Snips NLU [15].

4.2. Versión extendida con Snips NLU

Una vez validada la prueba de concepto con un documento, el siguiente paso ha consistido en desarrollar una versión extendida del chatbot, usando la herramienta Snips NLU, cumpliendo un par de requisitos:

- Contener información de varias Guías de Práctica Clínica, y alimentarse de forma automática y semiautomática a partir del texto extraído desde las guías, de acuerdo con el procedimiento explicado en los capítulos anteriores.

- Permitir la contextualización. Es frecuente en un chatbot, que una vez hecha una pregunta de forma muy explícita, las siguientes preguntas no sean tan explícitas en

cuanto a la información solicitada, porque está sobreentendida por el "contexto", es decir, por las preguntas y respuestas precedentes.

Para llevar a cabo esta versión extendida, primeramente se ha desarrollado una función en Python que, dado un documento en formato JSON, busca un patrón de pregunta en sus títulos, devolviendo una lista con todas las ocurrencias.

Para cada uno de estos títulos-pregunta, se ha diseñado otra función que devuelve todo el texto estándar contenido en ellos, es decir, la respuesta a las preguntas.

Automáticamente se crea un *intent* por documento, utilizando como frases de prueba todos los títulos-pregunta del mismo. Estos *intents* se guardan de forma persistente utilizando la sintaxis en YAML especificada por Snips NLU.

Tras crear los *intents* en YAML para todos los documentos, mediante la Command Line Interface que provee Snips NLU es posible generar un archivo JSON que se corresponde con el dataset propiamente dicho con el que se va a entrenar al motor del chatbot.

Con estos *intents* posteriormente se entrena un motor "general", que permite reconocer ante una pregunta cuál es el documento en el que es más probable que se encuentre la respuesta.

De forma paralela, también se crea un *intent* separado por cada pregunta de cada documento, además de almacenar en un fichero de texto la respuesta asociada a esa pregunta. Estos *intents* se usan para entrenar un motor por cada documento, además del "general"descrito anteriormente.

Los motores de clasificación son entrenados a partir de los dataset en JSON, y pueden ser guardados de forma persistente en un directorio.

La demo, de la cual un ejemplo se detalla en el Anexo A: Ejemplo de interacción con el chatbot, consiste en un script de Python que carga el motor "general" y pide en bucle frases para que sean introducidas por entrada estándar, de forma que una vez se reconoce cuál es el documento en cuestión al que se hace referencia, se carga el motor específico del documento para dar respuesta a esa pregunta y a las siguientes, de modo que en posteriores interacciones no es necesario volver a hacer referencia al problema de salud.

Capítulo 5

Metodología

En este apartado se describe la metodología de trabajo empleada en el proyecto, además de aspectos relativos a la planificación inicial de cada una de sus fases y una valoración sobre el ajuste de la realidad a esa planificación.

La metodología utilizada, debido a la propia naturaleza del proyecto, ha sido iterativa, añadiendo y modificando elementos del pipeline conforme se llevaba a cabo el desarrollo, con el objetivo de alcanzar el mejor resultado posible. Para llevar un control del código, se ha utilizado el sistema de control de versiones *Git*. En el repositorio GitHub del proyecto, se pueden observar cada uno de los cambios que se han ido realizando al software desarrollado.

En la planificación inicial del proyecto, se estimaron los siguientes plazos asociadas a sus respectivas tareas:

- Búsqueda de librerías y soluciones existentes para utilizar como punto de partida, análisis de las mismas y planteamiento de solución y desarrollo iterativo. Duración: 1 mes.

- Desarrollo iterativo del pipeline, contando con una fase de test al finalizar cada iteración. En esta fase se diseñan, prueban y depuran los distintos componentes que sea necesario desarrollar con el objetivo de obtener el resultado deseado. Duración: 2 meses.

- Redacción de la memoria. Duración: a lo largo de todo el proyecto, más 15 días a la finalización del desarrollo.

Para llevar un registro del tiempo dedicado al proyecto se ha hecho uso de la herramienta Clockify [16]. En ella, cada tiempo registrado se ha asociado con una tarea genérica. Las tareas establecidas han sido las siguientes:

- Desarrollo: Diseño y escritura del código del proyecto. En esta tarea se han

contabilizado tanto el pipeline como la prueba de concepto del chatbot (que no fue prevista en la planificación inicial).

- Pruebas: Fase de testing llevada a cabo después de cada pequeña iteración.
- Documentación: Investigación y búsqueda de información necesaria para el desarrollo del proyecto.
- Gestión: Principalmente anotación de resultados y redacción de la memoria, además de reuniones y comunicación con el tutor del proyecto.

En la gráfica de la Figura 5.1 se muestra el total de horas dedicadas, junto con los porcentajes de dedicación a cada una de las tareas genéricas descritas anteriormente.

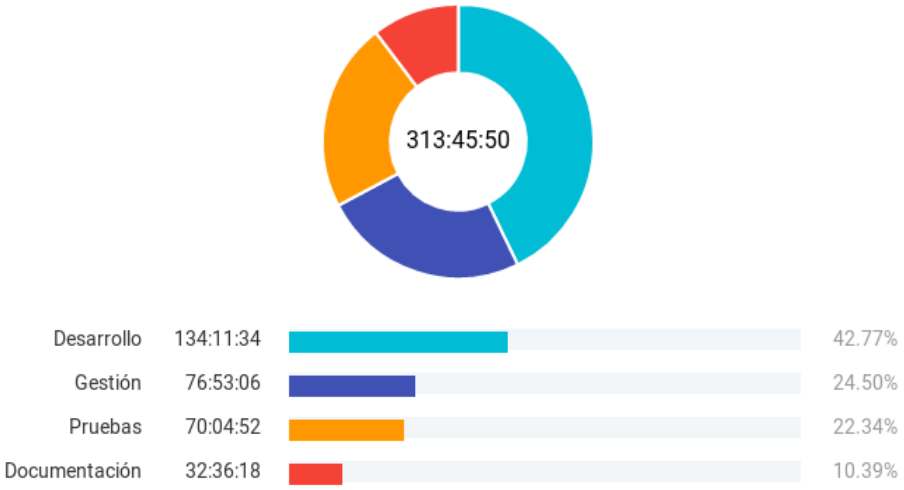


Figura 5.1: Tiempo dedicado a cada tarea genérica.

En la Figura 5.2 se puede comprobar como, de forma aproximada, las previsiones se ajustan a la realidad, teniendo en cuenta el reparto de las tareas a lo largo del tiempo de vida del proyecto.

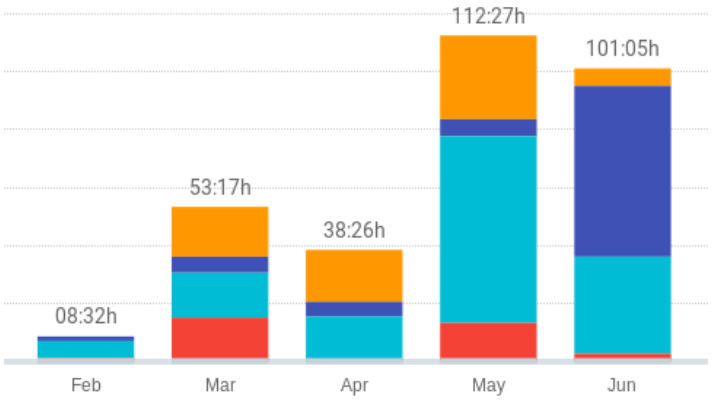


Figura 5.2: Registro de horas dedicadas a lo largo del tiempo de vida del proyecto.

Capítulo 6

Conclusiones

Una herramienta como la construida nunca podrá obtener resultados perfectos para cualquier PDF que se le proporcione como entrada, debido a las innumerables posibilidades que existen a la hora de construir un documento en este formato. El programa desarrollado se centra en obtener buenos resultados en Guías de Práctica Clínica; sin embargo, los resultados sobre el conjunto utilizado como prueba tampoco son perfectos. Siempre existe algún caso especial en alguno de los ficheros que provoca que el resultado no sea 100 % el deseado. Por otro lado, no es recomendable la búsqueda de la perfección en estos casos, ya que provoca, por una parte, una complejidad adicional en el procesamiento, y por otro lado un cierto "sobreajuste" a los documentos utilizados como conjunto de prueba (o entrenamiento), lo que puede suponer un mayor número de fallos en un posible uso en producción con otros documentos, frente a un diseño más genérico que consiga un equilibrio entre precisión y flexibilidad para un conjunto cualquiera de Guías de Práctica Clínica.

Por otro lado, el objetivo de la extracción de texto junto con su estructura a partir de documentos PDF que se ha realizado en este proyecto, no era obtener versiones perfectas reeditables de un documento; es decir, no se pretendía desarrollar un software capaz de transformar un documento PDF en un documento editable tipo "Microsoft Word". El objetivo era ser capaz de ir más allá de los extractores básicos de texto que utilizan los clásicos buscadores sobre texto plano (p.ej.: Lucene), para dotar de información adicional a los sistemas de Procesamiento de Lenguaje Natural, al permitirles contextualizar la información. Para este objetivo, el resultado del presente proyecto supone un avance considerable en comparación con los productos preexistentes que se han podido analizar.

Por último, cabe mencionar distintos trabajos a futuro que den continuidad a este Trabajo de Fin de Grado, ya sea en el ámbito de los chatbots, o sirviendo como punto de partida para otros casos de uso tales como buscadores especializados o sistemas más complejos de *query-answering* contextualizados. Futuras evoluciones del presente

trabajo, a través del procesamiento de Guías de Práctica Clínica o de cualquier otro tipo de documento científico médico, podrán permitir el desarrollo de sistemas de ayuda a la toma de decisiones clínicas, con base documental. Estos desarrollos forman parte de la línea de trabajo que la Secretaría de GuíaSalud, que gestiona el Instituto Aragonés de Ciencias de la Salud, tiene prevista dentro del proyecto de Biblioteca Inteligente GuiaSalud (BIGS), y dentro de la cual se integra este trabajo.

Capítulo 7

Bibliografía

[1] Instituto Aragonés de Ciencias de la Salud. Guía salud. https://portal.guiasalud.es/, 2020.

[2] Yusuke Shinyama. Pdfminer. https://pypi.org/project/pdfminer/.

[3] Pdfminer.six: Community manteined fork of pdfminer. https://github.com/pdfminer/pdfminer.six.

[4] Jeremy Singer-Vine. Plumb a pdf for detailed information about each char, rectangle, line, et cetera — and easily extract text and tables. https://github.com/jsvine/pdfplumber, 2020.

[5] Slate: The simplest way to extract text from pdfs in python. https://github.com/timClicks/slate, 2017.

[6] PDFlib GmbH. Tet features. https://www.pdflib.com/products/tet/features/, 2020.

[7] PDFlib GmbH. *PDFlib Text and Image Extraction Toolkit (TET) Manual*. 2019.

[8] Pypdf2 documentation. https://pythonhosted.org/PyPDF2/.

[9] Philipp Weissenbacher Rene Engelhard, Petr Mladek. soffice (1) - linux man pages. https://www.systutorials.com/docs/linux/man/1-soffice/#lbAH.

[10] Soren Boll Overgaard Gueorgui Ovtcharov, Rainer Dorsch. pdftohtml(1) [opendarwin man page]. https://www.unix.com/man-page/opendarwin/1/pdftohtml.

[11] Andrea Trevino. Introduction to k-means clustering - oracle data science. https://blogs.oracle.com/datascience/introduction-to-k-means-clustering, 2016.

[12] Mingzhou Song Haizhou Wang. Ckmeans.1d.dp: Optimal k-means clustering in one dimension by dynamic programming. *The R Journal*, 2011.

[13] Bill Mill. Ckmeans in python with numpy. `https://github.com/llimllib/ckmeans`, 2016.

[14] Dialogflow basics. `https://cloud.google.com/dialogflow/docs/basics`, 2020.

[15] Clement Doumouro Adrien Ball. Snips natural language understanding. `https://snips-nlu.readthedocs.io/en/latest/`, 2018.

[16] Clockify - 100 % free time tracking software. `https://clockify.me/`, 2020.

Lista de Figuras

1.1. Comparación de una misma página en formato PDF y en formato texto plano. 2

3.1. Esquema general del pipeline de procesamiento. 11
3.2. Comparación de una misma página antes y después de la extracción. . 13
3.3. Fragmentos HTML donde se indica la posición de inicio de página. . . . 14
3.4. Comparación de una misma página antes y después de eliminar elementos gráficos y pies de página. 15
3.5. Histograma de tamaños de fuente de un documento. 18
3.6. Referencia bibliográfica en uno de los documentos de prueba. 18
3.7. Comparación de una misma página antes y después de convertir de HTML a MarkDown. 21
3.8. Comparación de una misma página antes y después del procesamiento en formato MarkDown. 24
3.9. Nodo raíz del documento en formato JSON. 25
3.10. Nodo cualquiera del documento en formato JSON. 25

5.1. Tiempo dedicado a cada tarea genérica. 32
5.2. Registro de horas dedicadas a lo largo del tiempo de vida del proyecto. 33

A.1. Portada de GPC sobre Lactancia. 45
A.2. Portada de GPC sobre DHR. 46
A.3. Portada de GPC sobre Menopausia. 46
A.4. Portada de GPC sobre TDAH. 47
A.5. Portada de GPC sobre Depresión. 47
A.6. 1ª interacción: cómo se diagnostica el tdah. 48
A.7. Páginas del documento original donde se encuentra la información relativa a la 1ª interacción con el chatbot. 49
A.8. 2ª interacción: cuáles son los síntomas. 50

A.9. Páginas del documento original donde se encuentra la información relativa a la 2ª interacción con el chatbot. 51
A.10.3ª interacción: cómo se trata. 52
A.11.Página del documento original donde se encuentra la información relativa a la 3ª interacción con el chatbot. 52

B.1. Primera página del documento procesado. 54
B.2. Segunda página del documento procesado. 54
B.3. Documento en formato MarkDown. 55
B.4. Documento en formato JSON (1). 56
B.5. Documento en formato JSON (2). 57

Lista de Tablas

3.1. Estadísticas de ordenación de bloques en un conjunto de 40 documentos. 16

Anexos

Anexos A

Ejemplo de interacción con el chatbot

En este ejemplo de interacción se han utilizado 5 Guías de Práctica Clínica para entrenar al chatbot. Primero, se ha utilizado el pipeline de procesamiento para obtener una versión JSON estructurada de los documentos. Dicha versión en JSON ha sido utilizada para entrenar tanto el motor general de clasificación como los motores específicos de cada documento. Las Guías utilizadas, junto con una imagen de su primera página, se muestran en las siguientes 5 Figuras.

Figura A.1: Portada de GPC sobre Lactancia.

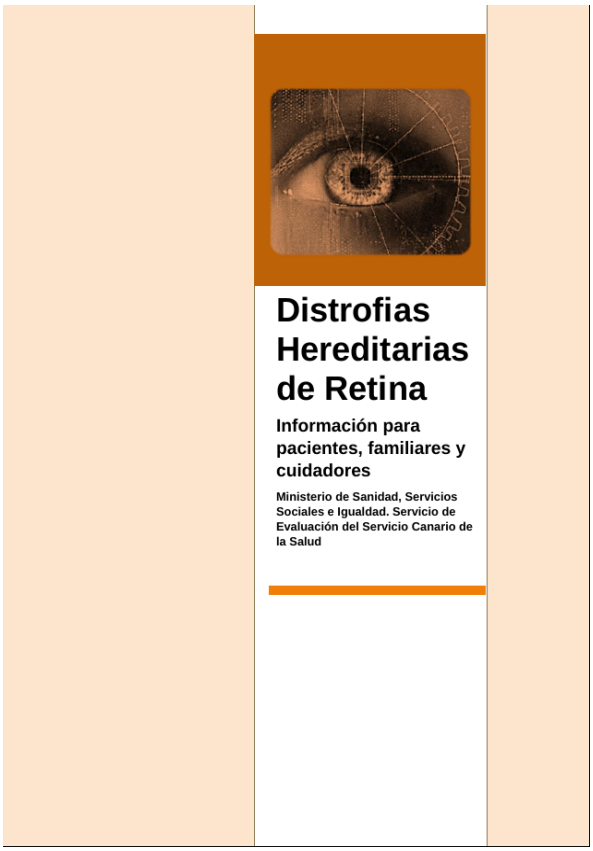


Figura A.2: Portada de GPC sobre DHR.

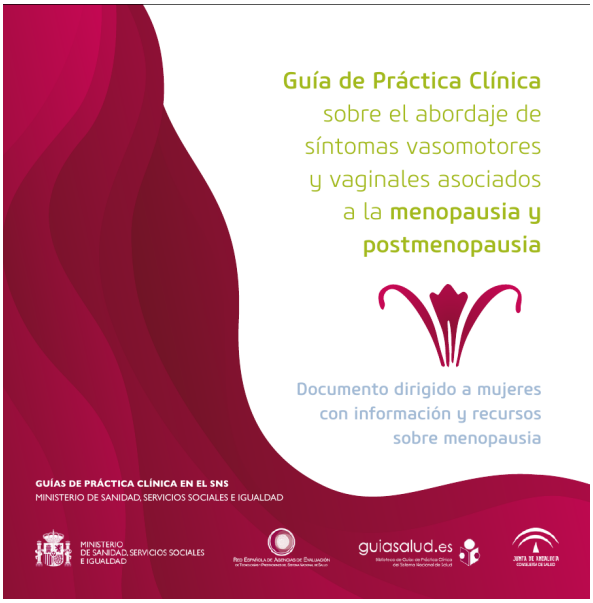


Figura A.3: Portada de GPC sobre Menopausia.

Figura A.4: Portada de GPC sobre TDAH.

Figura A.5: Portada de GPC sobre Depresión.

Como se ha mencionado en el capítulo correspondiente, el procesamiento de la primera *query* determina cuál es el documento más probable y, posteriormente, ofrece una respuesta utilizando el motor de clasificación específico del documento. La interacción propiamente dicha se produce a través de la introducción de frases en una terminal que funciona a modo de interfaz. La primera interacción y su resultado se puede observar en la Figura A.6, mientras que en la Figura A.7 se muestran las páginas originales del documento donde se encuentra la información.

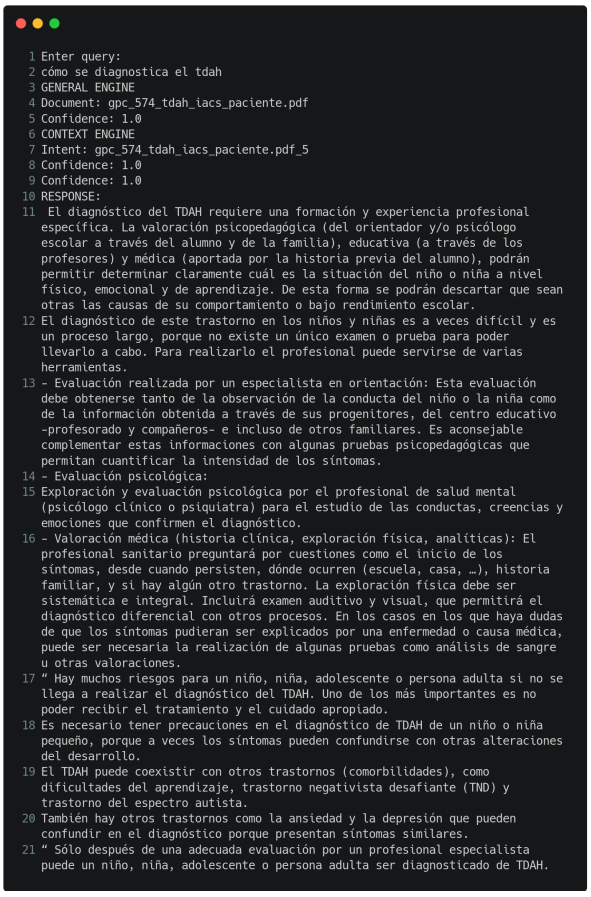


Figura A.6: 1ª interacción: cómo se diagnostica el tdah.

¿Cómo se diagnostica el TDAH?

El diagnóstico del TDAH requiere una formación y experiencia profesional específica. La valoración psicopedagógica (del orientador y/o psicólogo escolar a través del alumno y de la familia), educativa (a través de los profesores) y médica (aportada por la historia previa del alumno), podrán permitir determinar claramente cuál es la situación del niño o niña a nivel físico, emocional y de aprendizaje. De esta forma se podrán descartar que sean otras las causas de su comportamiento o bajo rendimiento escolar.

El diagnóstico de este trastorno en los niños y niñas es a veces difícil y es un proceso largo, porque no existe un único examen o prueba para poder llevarlo a cabo. Para realizarlo el profesional puede servirse de varias herramientas.

- *Evaluación realizada por un especialista en orientación*: Esta evaluación debe obtenerse tanto de la observación de la conducta del niño o la niña como de la información obtenida a través de sus progenitores, del centro educativo —profesorado y compañeros— e incluso de otros familiares. Es aconsejable complementar estas informaciones con algunas pruebas psicopedagógicas que permitan cuantificar la intensidad de los síntomas.
- *Evaluación psicológica:* Exploración y evaluación psicológica por el profesional de salud mental (psicólogo clínico o psiquiatra) para el estudio de las conductas, creencias y emociones que confirmen el diagnóstico.
- *Valoración médica (historia clínica, exploración física, analíticas)*: El profesional sanitario preguntará por cuestiones como el inicio de los síntomas, desde cuando persisten, dónde ocurren (escuela, casa, ...), historia familiar, y si hay algún otro trastorno. La exploración física debe ser sistemática e integral. Incluirá examen auditivo y visual, que permitirá el diagnóstico diferencial con otros procesos. En los casos en los que haya dudas de que los síntomas pudieran ser explicados por una enfermedad o causa médica, puede ser necesaria la realización de algunas pruebas como análisis de sangre u otras valoraciones.

Es necesario tener precauciones en el diagnóstico de TDAH de un niño o niña pequeño, porque a veces los síntomas pueden confundirse con otras alteraciones del desarrollo.

El TDAH puede coexistir con otros trastornos (comorbilidades), como dificultades del aprendizaje, trastorno negativista desafiante (TND) y trastorno del espectro autista. También hay otros trastornos como la ansiedad y la depresión que pueden confundir en el diagnóstico porque presentan síntomas similares.

> Hay muchos riesgos para un niño, niña, adolescente o persona adulta si no se llega a realizar el diagnóstico del TDAH. Uno de los más importantes es no poder recibir el tratamiento y el cuidado apropiado.

> Sólo después de una adecuada evaluación por un profesional especialista puede un niño, niña, adolescente o persona adulta ser diagnosticado de TDAH.

Figura A.7: Páginas del documento original donde se encuentra la información relativa a la 1ª interacción con el chatbot.

Después de la primera interacción con el chatbot, ya no es necesario volver a hacer referencia al problema de salud en la *query* introducida. Esto se puede comprobar en la Figura A.8 donde se muestra la segunda interacción con el chatbot.

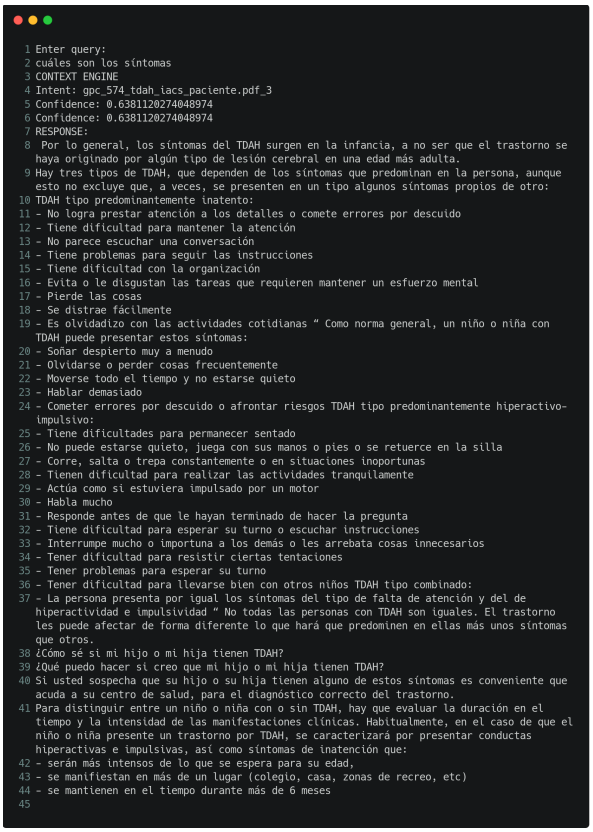
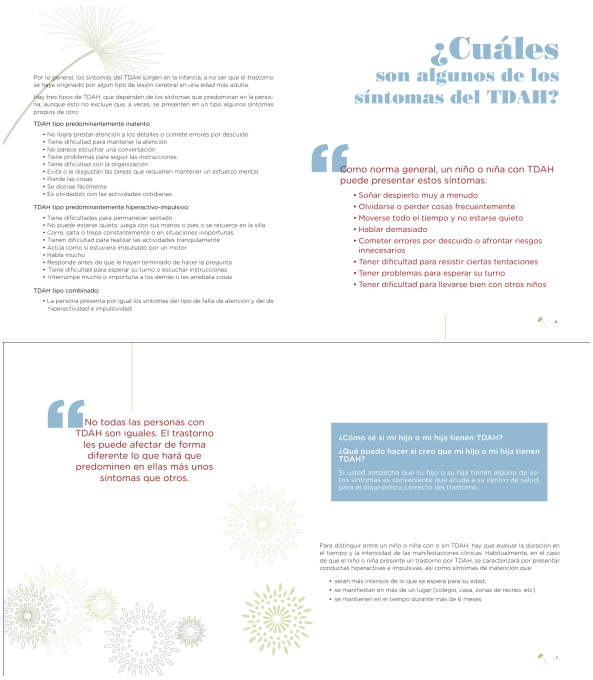


Figura A.8: 2ª interacción: cuáles son los síntomas.



Por lo general, los síntomas del TDAH surgen en la infancia, a no ser que el trastorno se haya originado por algún tipo de lesión cerebral en una edad más adulta.

Hay tres tipos de TDAH, que dependen de los síntomas que predominan en la persona, aunque esto no excluye que, a veces, se presenten en un tipo algunos síntomas propios de otro:

TDAH tipo predominantemente inatento:

- No logra prestar atención a los detalles o comete errores por descuido
- Tiene dificultad para mantener la atención
- No parece escuchar una conversación
- Tiene problemas para seguir las instrucciones
- Tiene dificultad con la organización
- Evita o le disgustan las tareas que requieren mantener un esfuerzo mental
- Pierde las cosas
- Se distrae fácilmente
- Es olvidadizo con las actividades cotidianas

TDAH tipo predominantemente hiperactivo-impulsivo:

- Tiene dificultades para permanecer sentado
- No puede estarse quieto, juega con sus manos o pies o se retuerce en la silla
- Corre, salta o trepa constantemente o en situaciones inoportunas
- Tienen dificultad para realizar las actividades tranquilamente
- Actúa como si estuviera impulsado por un motor
- Habla mucho
- Responde antes de que le hayan terminado de hacer la pregunta
- Tiene dificultad para esperar su turno o escuchar instrucciones
- Interrumpe mucho o importuna a los demás o les arrebata cosas

TDAH tipo combinado:

- La persona presenta por igual los síntomas del tipo de falta de atención y del de hiperactividad e impulsividad

¿Cuáles son algunos de los síntomas del TDAH?

Como norma general, un niño o niña con TDAH puede presentar estos síntomas:

- Soñar despierto muy a menudo
- Olvidarse o perder cosas frecuentemente
- Moverse todo el tiempo y no estarse quieto
- Hablar demasiado
- Cometer errores por descuido o afrontar riesgos innecesarios
- Tener dificultad para resistir ciertas tentaciones
- Tener problemas para esperar su turno
- Tener dificultad para llevarse bien con otros niños

4

No todas las personas con TDAH son iguales. El trastorno les puede afectar de forma diferente lo que hará que predominen en ellas más unos síntomas que otros.

¿Cómo sé si mi hijo o mi hija tienen TDAH?

¿Qué puedo hacer si creo que mi hijo o mi hija tienen TDAH?

Si usted sospecha que su hijo o su hija tienen alguno de estos síntomas es conveniente que acuda a su centro de salud, para el diagnóstico correcto del trastorno.

Para distinguir entre un niño o niña con o sin TDAH, hay que evaluar la duración en el tiempo y la intensidad de las manifestaciones clínicas. Habitualmente, en el caso de que el niño o niña presente un trastorno por TDAH, se caracterizará por presentar conductas hiperactivas e impulsivas, así como síntomas de inatención que:

- serán más intensos de lo que se espera para su edad,
- se manifiestan en más de un lugar (colegio, casa, zonas de recreo, etc)
- se mantienen en el tiempo durante más de 6 meses

7

Figura A.9: Páginas del documento original donde se encuentra la información relativa a la 2ª interacción con el chatbot.

La tercera y última interacción de este ejemplo se puede ver en la Figura A.10, a la par que la página original del documento donde se encuentra la información en la Figura A.11.

```
Enter query:
cómo se trata
CONTEXT ENGINE
Intent: gpc_574_tdah_iacs_paciente.pdf_6
Confidence: 0.6652406762573692
Confidence: 0.6652406762573692
RESPONSE:
 Aunque el tratamiento del TDAH no suele eliminar por completo los síntomas, si puede controlarlos. La mayoría de las personas requieren un tratamiento continuo para mantener sus síntomas bajo control.
Usted puede encontrar información sobre otros tratamientos de los que no hay estudios fiables sobre su eficacia. Cuando acuda a su centro de salud aporte información sobre cualquier sustancia, medicamento, producto de herbolario o medicina alternativa que esté tomando su hijo o hija para los problemas de TDAH.
Después de realizado el diagnóstico, los profesionales sanitarios y educativos son los indicados para preparar el plan de tratamiento más adecuado, dependiendo de las características propias de cada persona y de la familia y teniendo siempre en cuenta sus preferencias.
Hay que tener en cuenta la importancia que tiene conseguir el tratamiento apropiado para cada caso de TDAH, para eliminar las consecuencias negativas del propio trastorno.
“ Entre los tratamientos que resultan eficaces encontramos:
- las intervenciones psicológicas que incluyen 1- terapias cognitivo-conductuales (técnicas de control de la conducta y el pensamiento), 2- entrenamiento en habilidades sociales para las personas con TDAH, 3- intervención familiar de información y educación para padres, madres y cuidadores, El tratamiento debe adaptarse a las necesidades únicas de cada persona y su familia.
- las intervenciones psicopedagógicas, con programas especiales de apoyo en la escuela y
- las intervenciones farmacológicas.
Es lo que se llama un enfoque de “tratamiento multimodal”
```

Figura A.10: 3ª interacción: cómo se trata.

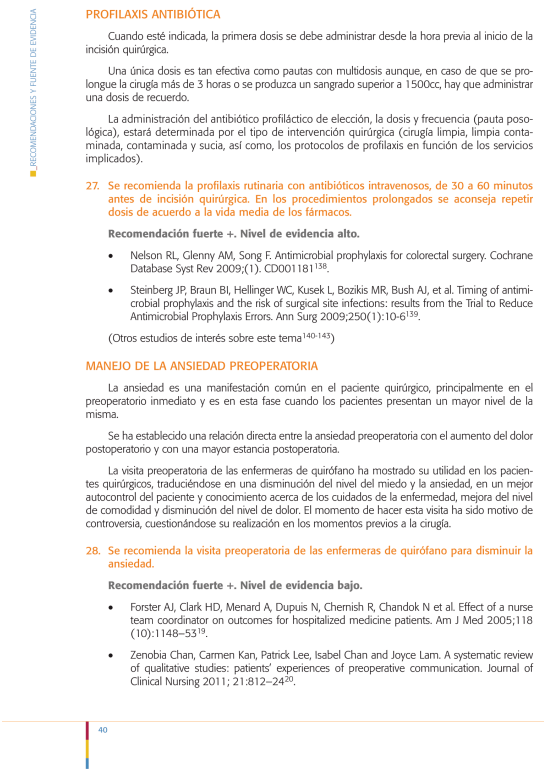
Figura A.11: Página del documento original donde se encuentra la información relativa a la 3ª interacción con el chatbot.

Anexos B

Ejemplo de extracción y procesado

Se han seleccionado 2 páginas de una Guía de Práctica Clínica para poder visualizar de manera más detallada los resultados que el software desarrollado puede ofrecer. En las Figuras B.1 y B.2 se muestran las páginas originales del documento, en la Figura B.3 la salida producida en formato MarkDown y en las Figuras B.4 y B.5 la salida en formato JSON.

Hay que tener en cuenta que el documento utilizado consta únicamente de estas dos páginas. Si se realizara el procesado de la Guía de Práctica Clínica al completo se podrían apreciar algunos cambios en este fragmento, como el nivel de relevancia de los títulos detectados.



PROFILAXIS ANTIBIÓTICA

Cuando esté indicada, la primera dosis se debe administrar desde la hora previa al inicio de la incisión quirúrgica.

Una única dosis es tan efectiva como pautas con multidosis aunque, en caso de que se prolongue la cirugía más de 3 horas o se produzca un sangrado superior a 1500cc, hay que administrar una dosis de recuerdo.

La administración del antibiótico profiláctico de elección, la dosis y frecuencia (pauta posológica), estará determinada por el tipo de intervención quirúrgica (cirugía limpia, limpia contaminada, contaminada y sucia, así como, los protocolos de profilaxis en función de los servicios implicados).

27. Se recomienda la profilaxis rutinaria con antibióticos intravenosos, de 30 a 60 minutos antes de incisión quirúrgica. En los procedimientos prolongados se aconseja repetir dosis de acuerdo a la vida media de los fármacos.

Recomendación fuerte +. Nivel de evidencia alto.

- Nelson RL, Glenny AM, Song F. Antimicrobial prophylaxis for colorectal surgery. Cochrane Database Syst Rev 2009;(1). CD001181[138].
- Steinberg JP, Braun BI, Hellinger WC, Kusek L, Bozikis MR, Bush AJ, et al. Timing of antimicrobial prophylaxis and the risk of surgical site infections: results from the Trial to Reduce Antimicrobial Prophylaxis Errors. Ann Surg 2009;250(1):10-6[139].

(Otros estudios de interés sobre este tema[140-143])

MANEJO DE LA ANSIEDAD PREOPERATORIA

La ansiedad es una manifestación común en el paciente quirúrgico, principalmente en el preoperatorio inmediato y es en esta fase cuando los pacientes presentan un mayor nivel de la misma.

Se ha establecido una relación directa entre la ansiedad preoperatoria con el aumento del dolor postoperatorio y con una mayor estancia postoperatoria.

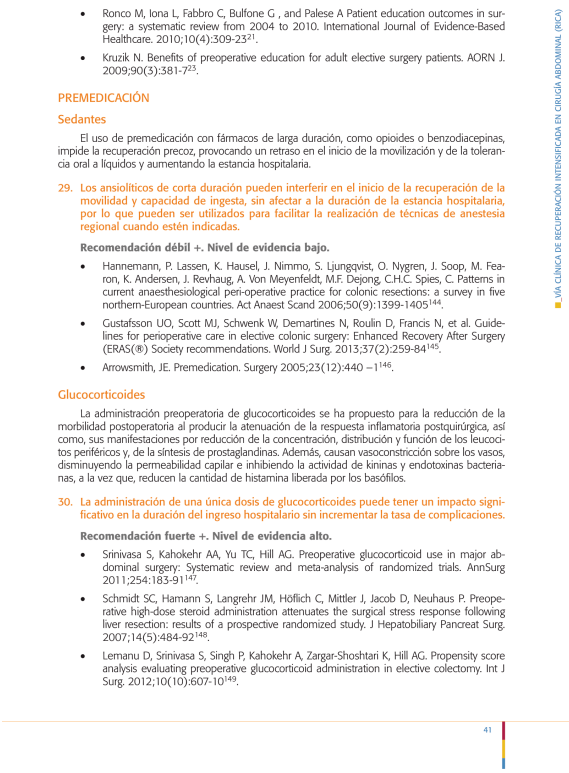
La visita preoperatoria de las enfermeras de quirófano ha mostrado su utilidad en los pacientes quirúrgicos, traduciéndose en una disminución del nivel del miedo y la ansiedad, en un mejor autocontrol del paciente y conocimiento acerca de los cuidados de la enfermedad, mejora del nivel de comodidad y disminución del nivel de dolor. El momento de hacer esta visita ha sido motivo de controversia, cuestionándose su realización en los momentos previos a la cirugía.

28. Se recomienda la visita preoperatoria de las enfermeras de quirófano para disminuir la ansiedad.

Recomendación fuerte +. Nivel de evidencia bajo.

- Forster AJ, Clark HD, Menard A, Dupuis N, Chernish R, Chandok N et al. Effect of a nurse team coordinator on outcomes for hospitalized medicine patients. Am J Med 2005;118 (10):1148–53[19].
- Zenobia Chan, Carmen Kan, Patrick Lee, Isabel Chan and Joyce Lam. A systematic review of qualitative studies: patients' experiences of preoperative communication. Journal of Clinical Nursing 2011; 21:812–24[20].

Figura B.1: Primera página del documento procesado.



- Ronco M, Iona L, Fabbro C, Bulfone G , and Palese A Patient education outcomes in surgery: a systematic review from 2004 to 2010. International Journal of Evidence-Based Healthcare. 2010;10(4):309-23[21].
- Kruzik N. Benefits of preoperative education for adult elective surgery patients. AORN J. 2009;90(3):381-7[23].

PREMEDICACIÓN

Sedantes

El uso de premedicación con fármacos de larga duración, como opioides o benzodiacepinas, impide la recuperación precoz, provocando un retraso en el inicio de la movilización y de la tolerancia oral a líquidos y aumentando la estancia hospitalaria.

29. Los ansiolíticos de corta duración pueden interferir en el inicio de la recuperación de la movilidad y capacidad de ingesta, sin afectar a la duración de la estancia hospitalaria, por lo que pueden ser utilizados para facilitar la realización de técnicas de anestesia regional cuando estén indicadas.

Recomendación débil +. Nivel de evidencia bajo.

- Hannemann, P. Lassen, K. Hausel, J. Nimmo, S. Ljungqvist, O. Nygren, J. Soop, M. Fearon, K. Andersen, J. Revhaug, A. Von Meyenfeldt, M.F. Dejong, C.H.C. Spies, C. Patterns in current anaesthesiological peri-operative practice for colonic resections: a survey in five northern-European countries. Act Anaest Scand 2006;50(9):1399-1405[144].
- Gustafsson UO, Scott MJ, Schwenk W, Demartines N, Roulin D, Francis N, et al. Guidelines for perioperative care in elective colonic surgery: Enhanced Recovery After Surgery (ERAS(®) Society recommendations. World J Surg. 2013;37(2):259-84[145].
- Arrowsmith, JE. Premedication. Surgery 2005;23(12):440 –1[146].

Glucocorticoides

La administración preoperatoria de glucocorticoides se ha propuesto para la reducción de la morbilidad postoperatoria al producir la atenuación de la respuesta inflamatoria postquirúrgica, así como, sus manifestaciones por reducción de la concentración, distribución y función de los leucocitos periféricos y, de la síntesis de prostaglandinas. Además, causan vasoconstricción sobre los vasos, disminuyendo la permeabilidad capilar e inhibiendo la actividad de kininas y endotoxinas bacterianas, a la vez que, reducen la cantidad de histamina liberada por los basófilos.

30. La administración de una única dosis de glucocorticoides puede tener un impacto significativo en la duración del ingreso hospitalario sin incrementar la tasa de complicaciones.

Recomendación fuerte +. Nivel de evidencia alto.

- Srinivasa S, Kahokehr AA, Yu TC, Hill AG. Preoperative glucocorticoid use in major abdominal surgery: Systematic review and meta-analysis of randomized trials. AnnSurg 2011;254:183-91[147].
- Schmidt SC, Hamann S, Langrehr JM, Höflich C, Mittler J, Jacob D, Neuhaus P. Preoperative high-dose steroid administration attenuates the surgical stress response following liver resection: results of a prospective randomized study. J Hepatobiliary Pancreat Surg. 2007;14(5):484-92[148].
- Lemanu D, Srinivasa S, Singh P, Kahokehr A, Zargar-Shoshtari K, Hill AG. Propensity score analysis evaluating preoperative glucocorticoid administration in elective colectomy. Int J Surg. 2012;10(10):607-10[149].

Figura B.2: Segunda página del documento procesado.

```
# PROFILAXIS ANTIBIÓTICA
Cuando esté indicada, la primera dosis se debe administrar desde la hora previa al inicio de la incisión quirúrgica.

Una única dosis es tan efectiva como pautas con multidosis aunque, en caso de que se prolongue la cirugía más de 3 horas o se produzca un sangrado superior a 1500cc, hay que administrar una dosis de recuerdo.

La administración del antibiótico profiláctico de elección, la dosis y frecuencia (pauta posológica), estará determinada por el tipo de intervención quirúrgica (cirugía limpia, limpia contaminada, contaminada y sucia, así como, los protocolos de profilaxis en función de los servicios implicados).

27. Se recomienda la profilaxis rutinaria con antibióticos intravenosos, de 30 a 60 minutos antes de incisión quirúrgica. En los procedimientos prolongados se aconseja repetir dosis de acuerdo a la vida media de los fármacos.

Recomendación fuerte +. Nivel de evidencia alto.

- Nelson RL, Glenny AM, Song F. Antimicrobial prophylaxis for colorectal surgery. Cochrane Database Syst Rev 2009;(1). CD001181.

Steinberg JP, Braun BI, Hellinger WC, Kusek L, Bozikis MR, Bush AJ, et al. Timing of antimicrobial prophylaxis and the risk of surgical site infections: results from the Trial to Reduce Antimicrobial Prophylaxis Errors. Ann Surg 2009;250(1):10-6.

- (Otros estudios de interés sobre este tema )

# MANEJO DE LA ANSIEDAD PREOPERATORIA
La ansiedad es una manifestación común en el paciente quirúrgico, principalmente en el preoperatorio inmediato y es en esta fase cuando los pacientes presentan un mayor nivel de la misma.

Se ha establecido una relación directa entre la ansiedad preoperatoria con el aumento del dolor postoperatorio y con una mayor estancia postoperatoria.

La visita preoperatoria de las enfermeras de quirófano ha mostrado su utilidad en los pacientes quirúrgicos, traduciéndose en una disminución del nivel del miedo y la ansiedad, en un mejor autocontrol del paciente y conocimiento acerca de los cuidados de la enfermedad, mejora del nivel de comodidad y disminución del nivel de dolor. El momento de hacer esta visita ha sido motivo de controversia, cuestionándose su realización en los momentos previos a la cirugía.

28. Se recomienda la visita preoperatoria de las enfermeras de quirófano para disminuir la ansiedad.

Recomendación fuerte +. Nivel de evidencia bajo.

Forster AJ, Clark HD, Menard A, Dupuis N, Chernish R, Chandok N et al. Effect of a nurse team coordinator on outcomes for hospitalized medicine patients. Am J Med 2005;118 (10):1148-53.

- Zenobia Chan, Carmen Kan, Patrick Lee, Isabel Chan and Joyce Lam. A systematic review of qualitative studies: patients' experiences of preoperative communication. Journal of Clinical Nursing 2011; 21:812-24.

- Ronco M, Iona L, Fabbro C, Bulfone G, and Palese A Patient education outcomes in surgery: a systematic review from 2004 to 2010. International Journal of Evidence-Based Healthcare. 2010;10(4):309-23.

- Kruzik N. Benefits of preoperative education for adult elective surgery patients. AORN J.
2009;90(3):381-7.

# PREMEDICACIÓN Sedantes
El uso de premedicación con fármacos de larga duración, como opioides o benzodiacepinas, impide la recuperación precoz, provocando un retraso en el inicio de la movilización y de la tolerancia oral a líquidos y aumentando la estancia hospitalaria.

29. Los ansiolíticos de corta duración pueden interferir en el inicio de la recuperación de la movilidad y capacidad de ingesta, sin afectar a la duración de la estancia hospitalaria, por lo que pueden ser utilizados para facilitar la realización de técnicas de anestesia regional cuando estén indicadas.

Recomendación débil +. Nivel de evidencia bajo.

- Hannemann, P. Lassen, K. Hausel, J. Nimmo, S. Ljungqvist, O. Nygren, J. Soop, M. Fearon, K. Andersen, J. Revhaug, A. Von Meyenfeldt, M.F. Dejong, C.H.C. Spies, C. Patterns in current anaesthesiological peri-operative practice for colonic resections: a survey in five northern-European countries. Act Anaest Scand 2006;50(9):1399-1405.

- Gustafsson UO, Scott MJ, Schwenk W, Demartines N, Roulin D, Francis N, et al. Guidelines for perioperative care in elective colonic surgery: Enhanced Recovery After Surgery (ERAS(®) Society recommendations. World J Surg. 2013;37(2):259-84.

Arrowsmith, JE. Premedication. Surgery 2005;23(12):440 -1.

# Glucocorticoides
La administración preoperatoria de glucocorticoides se ha propuesto para la reducción de la morbilidad postoperatoria al producir la atenuación de la respuesta inflamatoria postquirúrgica, así como, sus manifestaciones por reducción de la concentración, distribución y función de los leucocitos periféricos y, de la síntesis de prostaglandinas. Además, causan vasoconstricción sobre los vasos, disminuyendo la permeabilidad capilar e inhibiendo la actividad de kininas y endotoxinas bacterianas, a la vez que, reducen la cantidad de histamina liberada por los basófilos.

30. La administración de una única dosis de glucocorticoides puede tener un impacto significativo en la duración del ingreso hospitalario sin incrementar la tasa de complicaciones.

Recomendación fuerte +. Nivel de evidencia alto.

Srinivasa S, Kahokehr AA, Yu TC, Hill AG. Preoperative glucocorticoid use in major abdominal surgery: Systematic review and meta-analysis of randomized trials. AnnSurg 2011;254:183-91.

- Schmidt SC, Hamann S, Langrehr JM, Höflich C, Mittler J, Jacob D, Neuhaus P. Preoperative high-dose steroid administration attenuates the surgical stress response following liver resection: results of a prospective randomized study. J Hepatobiliary Pancreat Surg.

2007;14(5):484-92.

- Lemanu D, Srinivasa S, Singh P, Kahokehr A, Zargar-Shoshtari K, Hill AG. Propensity score analysis evaluating preoperative glucocorticoid administration in elective colectomy. Int J Surg. 2012;10(10):607-10.
```

Figura B.3: Documento en formato MarkDown.

```
{
    "document": "fragmento_rica.pdf",
    "level": 0,
    "content": [
        {
            "text": "PROFILAXIS ANTIBIÓTICA",
            "level": 1,
            "content": [
                {
                    "text": "Cuando esté indicada, la primera dosis se debe administrar desde la hora previa al inicio de la incisión quirúrgica.",
                    "level": 7
                },
                {
                    "text": "Una única dosis es tan efectiva como pautas con multidosis aunque, en caso de que se prolongue la cirugía más de 3 horas o se produzca un sangrado superior a 1500cc, hay que administrar una dosis de recuerdo.",
                    "level": 7
                },
                {
                    "text": "La administración del antibiótico profiláctico de elección, la dosis y frecuencia (pauta posológica), estará determinada por el tipo de intervención quirúrgica (cirugía limpia, limpia contaminada, contaminada y sucia, así como, los protocolos de profilaxis en función de los servicios implicados).",
                    "level": 7
                },
                {
                    "text": "27. Se recomienda la profilaxis rutinaria con antibióticos intravenosos, de 30 a 60 minutos antes de incisión quirúrgica. En los procedimientos prolongados se aconseja repetir dosis de acuerdo a la vida media de los fármacos.",
                    "level": 7
                },
                {
                    "text": "Recomendación fuerte +. Nivel de evidencia alto.",
                    "level": 7
                },
                {
                    "text": "- Nelson RL, Glenny AM, Song F. Antimicrobial prophylaxis for colorectal surgery. Cochrane Database Syst Rev 2009;(1). CD001181.",
                    "level": 7
                },
                {
                    "text": "Steinberg JP, Braun BI, Hellinger WC, Kusek L, Bozikis MR, Bush AJ, et al. Timing of antimicrobial prophylaxis and the risk of surgical site infections: results from the Trial to Reduce Antimicrobial Prophylaxis Errors. Ann Surg 2009;250(1):10-6.",
                    "level": 7
                },
                {
                    "text": "- (Otros estudios de interés sobre este tema )",
                    "level": 7
                }
            ]
        },
        {
            "text": "MANEJO DE LA ANSIEDAD PREOPERATORIA",
            "level": 1,
            "content": [
                {
                    "text": "La ansiedad es una manifestación común en el paciente quirúrgico, principalmente en el preoperatorio inmediato y es en esta fase cuando los pacientes presentan un mayor nivel de la misma.",
                    "level": 7
                },
                {
                    "text": "Se ha establecido una relación directa entre la ansiedad preoperatoria con el aumento del dolor postoperatorio y con una mayor estancia postoperatoria.",
                    "level": 7
                },
                {
                    "text": "La visita preoperatoria de las enfermeras de quirófano ha mostrado su utilidad en los pacientes quirúrgicos, traduciéndose en una disminución del nivel del miedo y la ansiedad, en un mejor autocontrol del paciente y conocimiento acerca de los cuidados de la enfermedad, mejora del nivel de comodidad y disminución del nivel de dolor. El momento de hacer esta visita ha sido motivo de controversia, cuestionándose su realización en los momentos previos a la cirugía.",
                    "level": 7
                },
                {
                    "text": "28. Se recomienda la visita preoperatoria de las enfermeras de quirófano para disminuir la ansiedad.",
                    "level": 7
                },
                {
                    "text": "Recomendación fuerte +. Nivel de evidencia bajo.",
                    "level": 7
                },
                {
                    "text": "Forster AJ, Clark HD, Menard A, Dupuis N, Chernish R, Chandok N et al. Effect of a nurse team coordinator on outcomes for hospitalized medicine patients. Am J Med 2005;118 (10):1148-53.",
                    "level": 7
                },
                {
                    "text": "- Zenobia Chan, Carmen Kan, Patrick Lee, Isabel Chan and Joyce Lam. A systematic review of qualitative studies: patients' experiences of preoperative communication. Journal of Clinical Nursing 2011; 21:812-24.",
                    "level": 7
                },
                {
                    "text": "- Ronco M, Iona L, Fabbro C, Bulfone G, and Palese A Patient education outcomes in surgery: a systematic review from 2004 to 2010. International Journal of Evidence-Based Healthcare. 2010;10(4):309-23.",
                    "level": 7
                },
                {
                    "text": "- Kruzik N. Benefits of preoperative education for adult elective surgery patients. AORN J.",
                    "level": 7
                }
```


Figura B.4: Documento en formato JSON (1).

```
                {
                    "text": "2009;90(3):381-7.",
                    "level": 7
                }
            ]
        },
        {
            "text": "PREMEDICACIÓN Sedantes",
            "level": 1,
            "content": [
                {
                    "text": "El uso de premedicación con fármacos de larga duración, como opioides o benzodiacepinas, impide la recuperación precoz, provocando un retraso en el inicio de la movilización y de la tolerancia oral a líquidos y aumentando la estancia hospitalaria.",
                    "level": 7
                },
                {
                    "text": "29. Los ansiolíticos de corta duración pueden interferir en el inicio de la recuperación de la movilidad y capacidad de ingesta, sin afectar a la duración de la estancia hospitalaria, por lo que pueden ser utilizados para facilitar la realización de técnicas de anestesia regional cuando estén indicadas.",
                    "level": 7
                },
                {
                    "text": "Recomendación débil +. Nivel de evidencia bajo.",
                    "level": 7
                },
                {
                    "text": "- Hannemann, P. Lassen, K. Hausel, J. Nimmo, S. Ljungqvist, O. Nygren, J. Soop, M. Fearon, K. Andersen, J. Revhaug, A. Von Meyenfeldt, M.F. Dejong, C.H.C. Spies, C. Patterns in current anaesthesiological peri-operative practice for colonic resections: a survey in five northern-European countries. Act Anaest Scand 2006;50(9):1399-1405.",
                    "level": 7
                },
                {
                    "text": "- Gustafsson UO, Scott MJ, Schwenk W, Demartines N, Roulin D, Francis N, et al. Guidelines for perioperative care in elective colonic surgery: Enhanced Recovery After Surgery (ERAS(®) Society recommendations. World J Surg. 2013;37(2):259-84.",
                    "level": 7
                },
                {
                    "text": "Arrowsmith, JE. Premedication. Surgery 2005;23(12):440 -1.",
                    "level": 7
                }
            ]
        },
        {
            "text": "Glucocorticoides",
            "level": 1,
            "content": [
                {
                    "text": "La administración preoperatoria de glucocorticoides se ha propuesto para la reducción de la morbilidad postoperatoria al producir la atenuación de la respuesta inflamatoria postquirúrgica, así como, sus manifestaciones por reducción de la concentración, distribución y función de los leucocitos periféricos y, de la síntesis de prostaglandinas. Además, causan vasoconstricción sobre los vasos, disminuyendo la permeabilidad capilar e inhibiendo la actividad de kininas y endotoxinas bacterianas, a la vez que, reducen la cantidad de histamina liberada por los basófilos.",
                    "level": 7
                },
                {
                    "text": "30. La administración de una única dosis de glucocorticoides puede tener un impacto significativo en la duración del ingreso hospitalario sin incrementar la tasa de complicaciones.",
                    "level": 7
                },
                {
                    "text": "Recomendación fuerte +. Nivel de evidencia alto.",
                    "level": 7
                },
                {
                    "text": "Srinivasa S, Kahokehr AA, Yu TC, Hill AG. Preoperative glucocorticoid use in major abdominal surgery: Systematic review and meta-analysis of randomized trials. AnnSurg 2011;254:183-91.",
                    "level": 7
                },
                {
                    "text": "- Schmidt SC, Hamann S, Langrehr JM, Höflich C, Mittler J, Jacob D, Neuhaus P. Preoperative high-dose steroid administration attenuates the surgical stress response following liver resection: results of a prospective randomized study. J Hepatobiliary Pancreat Surg.",
                    "level": 7
                },
                {
                    "text": "2007;14(5):484-92.",
                    "level": 7
                },
                {
                    "text": "- Lemanu D, Srinivasa S, Singh P, Kahokehr A, Zargar-Shoshtari K, Hill AG. Propensity score analysis evaluating preoperative glucocorticoid administration in elective colectomy. Int J Surg. 2012;10(10):607-10.",
                    "level": 7
                }
            ]
        }
    ]
}
```


Figura B.5: Documento en formato JSON (2).